Semantic Video Search*

A.W.M. Smeulders, J.C. van Gemert, B. Huurnink, D.C. Koelma, O. de Rooij,
K.E.A. van de Sande, C.G.M. Snoek, C.J. Veenman, M. Worring
Intelligent Systems Lab Amsterdam, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
http://www.mediamill.nl

Abstract

In this paper we describe the current performance of our MediaMill system as presented in the TRECVID 2006 benchmark for video search engines. The MediaMill team participated in two tasks: concept detection and search. For concept detection we use the MediaMill Challenge as experimental platform. The MediaMill Challenge divides the generic video indexing problem into a visual-only, textual-only, early fusion, late fusion, and combined analysis experiment. We provide a baseline implementation for each experiment together with baseline results. We extract image features, on global, regional, and keypoint level, which we combine with various supervised learners. A late fusion approach of visual-only analysis methods using geometric mean was our most successful run. With this run we conquer the Challenge baseline by more than 50%. Our concept detection experiments have resulted in the best score for three concepts: i.e. desert*,* flag us*, and* charts*. What is more, using LSCOM annotations, our visual-only approach generalizes well to a set of 491 concept detectors. To handle such a large thesaurus in retrieval, an engine is developed which allows users to select relevant concept detectors based on interactive browsing using advanced visualizations. Similar to previous years our best interactive search runs yield top performance, ranking 2nd and 6th overall.*

1 Introduction

Video is quickly becoming the most important information carrier of our time. Not only is the amount of bits on the Internet spend on video already larger than any of the other information carriers, video is also growing fastest in the items. The newest wave of successful Internet companies rely on video. One may observe that the preference for video was only hindered by the absence of digital recorders, sufficient storage space and high-speed networks and computers. Once all components are there, the audio-visual format immediately resumes its premiere place in communication. This puts the pressure on engines providing access to video information. Ten DVDs of home videos, hundreds of professional video tapes, or thousands of video tapes in a broadcast archive cannot be disclosed without annotations: one option is manual annotation, another one is social tagging via the Internet, and a third one is automatic (interactive) annotation. In the reality of the future for all three classes of video repositories, all three different techniques will appear useful (in combination).

Most commercial video search engines such as Google, Blinkx, and YouTube provide access to their repositories 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, or a transcript. This results in disappointing performance when the visual content is not reflected in the associated text. In addition, when the videos originate from non-English speaking countries, such as China, Lebanon, or the Netherlands, querying the content becomes even harder as automatic speech recognition results are so much poorer. Additional visual analysis yields more robustness. Thus, in video retrieval a recent trend is to learn a lexicon of semantic concepts from multimedia examples and to employ these as entry points in querying the collection.

Previously we presented the *MediaMill 2005* semantic video search engine [20] using a 101 concept lexicon. For our current system we made a jump to a thesaurus of 491 concepts. The items vary from pure format like a detected *split screen*, or a style like an *interview*, or an object like a *horse*, or an event like an *airplane take off*. Any one of those brings an understanding of the current content. The elements in such a thesaurus offer users a semantic entry to video by allowing them to query on presence or absence of content elements.

*An extended version of this paper appeared in the TRECVID 2006 workshop proceedings as *The MediaMill TRECVID 2006 Semantic Video Search Engine* [21]

For a user, however, selecting the right topic from the large thesaurus is difficult. We therefore developed an interactive search engine with two novel browsers that present retrieval results using advanced visualizations. Taken together, the *MediaMill 2006* semantic video search engine provides users with semantic access to news video archives.

The remainder of the paper is organized as follows. We first define our semantic video indexing architecture in Section 2, introducing the MediaMil Challenge and our mostly visual analysis approach for this year's TRECVID. Then we highlight our semantic video retrieval engine in Section 3, which includes novel video browsers.

2 Semantic Video Indexing

Our generic semantic video indexing architecture is based on the semantic pathfinder [20, 22]. It is founded on the observation that produced video is the result of an authoring process. The semantic pathfinder selects the best path through content analysis, style analysis, and context analysis. This year we use a semantic pathfinder that relies mainly on (visual) content analysis, where the MediaMill Challenge [24] replaces the content analysis step. In this section we will highlight which components and experiments of the Challenge have been replaced by more elaborate analysis, learning, and combination schemes.

2.1 MediaMill Challenge

TRECVID has been of pivotal importance in assessing complete video indexing methods on their relative merit. In the course of the TRECVID benchmark some groups have shared annotations, like LSCOM [15], donated features, like the camera shot segmentation by CLIPS-IMAG [18], speech recognition results donated by LIMSI [6] and various multimedia features donated by Informedia [26]. In addition, all participants share their results on common test data for a limited lexicon of typically 10 high-level concepts. Until recently, however, nobody has provided low-level features and detected semantic concepts for a large lexicon on both training and test data, while these are crucial assets for repeatability of intermediate analysis steps.

The MediaMill Challenge [24] divides the generic video indexing problem into a visual-only, textual-only, early fusion, late fusion, and combined analysis experiment, see Fig.1. We provide a baseline implementation for each experiment together with baseline results for a lexicon containing 101 semantic concept detectors. The 85 hours of training data from the TRECVID 2005 corpus forms the basis for the MediaMill Challenge. We divided this archive a priori into a non-overlapping train and test set. The Challenge train set $\mathcal{A}$ contains 70% of the data, and the Challenge test set $\mathcal{B}$ holds the remaining 30%. The Challenge

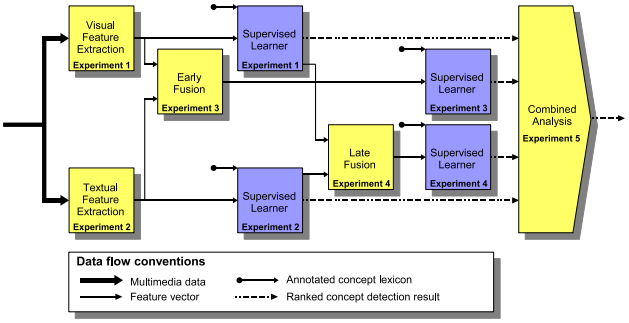

Figure 1: Data flow within the MediaMill Challenge for generic video indexing of 101 semantic concepts [24]. Experiment 1 and 2 focus on unimodal analysis, yielding a visual and a textual concept classification. Experiment 3 and 4 employ an early and late fusion scheme respectively. The Challenge allows for the construction of four classifiers for each concept. In experiment 5, an optimum is selected based on combined analysis.

package has been used by several teams for their 2006 system, either for comparison or as a building block for their submission.

2.2 Supervised Learners

We perceive concept detection in video as a pattern recognition problem. Given pattern $\vec{x}$, part of a shot i, the aim is to obtain a probability measure, which indicates whether semantic concept ω_j is present in shot i. Similar to the MediaMill Challenge, we use the Support Vector Machine (SVM) framework [25] for supervised learning of concepts. Here we use the LIBSVM implementation [2] with radial basis function and probabilistic output [16]. We obtain good SVM parameter settings by using an iterative search on a large number of SVM parameter combinations. The MediaMill Challenge optimizes SVM parameters that aim to balance positive and negative examples (w_{+1} and w_{-1}). Here we take the γ parameter into account also. We measure average precision performance of all parameter combinations and select the combination that yields the best performance. We use a 3-fold cross validation on Challenge train set $\mathcal{A}$ to prevent overfitting of parameters. Rather than using regular cross-validation for SVM parameter optimization, we also experiment with the recently proposed *episode-constrained* cross-validation method, as this method is known to yield a more accurate estimate of classifier performance [8].

In addition to the SVM we also experiment with logistic regression and Fisher's linear discriminant [4]. While both classifiers are known to be less effective than SVM, in terms of concept detection performance, they require no parameter tuning so classification is relatively cheap. Logistic regression performs a maximum likelihood estimation of

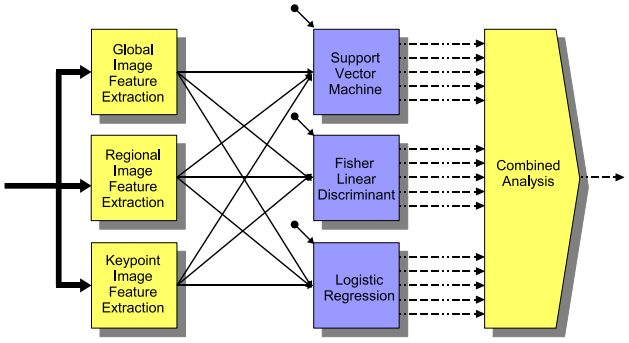

Figure 2: Simplified overview of our visual-only analysis approach for TRECVID 2006, using the conventions of Fig.1.

weights for the different feature dimensions, under the assumption that the observed training data was generated by a binomial model. In contrast, the Fisher's linear discriminant assumes normal distribution. It is used to find the linear combination of features which best separates two classes. It minimizes the errors in the least square sense. We use the resulting combinations as a linear classifier. For both classifiers we use the PRTools implementation [3]. All three classifiers yield a probability measure $p(\omega_j|\vec{x}_i)$, which we use to rank and to combine concept detection results.

2.3 Visual-Only Analysis

Given the promising performance of our visual features in last years benchmark, we have concentrated this years' efforts mainly on visual-only analysis, i.e. experiment 1 of the MediaMill Challenge. We extract image features on three levels of abstraction, namely: global level, region level, and keypoint level. All visual features are used in isolation or in combination, with the three supervised learners. Finally, we combine the individual concept detectors in several ways and select the combination that maximizes validation set performance.

2.3.1 Global Image Feature Extraction

We rely on Wiccest features for global image feature extraction. Wiccest features [10] utilize natural image statistics to effectively model texture information. Texture is described by the distribution of edges in a certain image. Hence, a histogram of a Gaussian derivative filter is used to represent the edge statistics. The complete range of image statistics in natural textures can be well modeled with an integrated Weibull distribution [9]. This distribution is given by

$$f(r) = \frac{\gamma}{2\gamma^{\frac{1}{\gamma}}\beta\Gamma(\frac{1}{\gamma})} \exp\left\{-\frac{1}{\gamma}\left|\frac{r-\mu}{\beta}\right|^{\gamma}\right\}, \qquad (1)$$

where r is the edge response to the Gaussian derivative filter and $\Gamma(\cdot)$ is the complete Gamma function, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-1}dt$. The parameter β denotes the width of the distribution, the parameter γ represents the 'peakness' of the distribution, and the parameter μ denotes the mode of the distribution. The position of the mode is influenced by uneven illumination and colored illumination. Hence, to achieve color constancy the values for μ is ignored.

The integrated Weibull distribution can be estimated from a histogram of filter responses with a maximum likelihood estimator as described in [10]. The parameters μ, β and γ are estimated by taking the derivatives of the integrated Weibull distribution to the respective parameters and setting them to zero.

2.3.2 Regional Image Feature Extraction

We also use Wiccest features for regional image feature extraction. We divide an input frame into multiple overlapping regions, and compute for each region the similarity to 15 proto-concepts [7].

In addition to the Wiccest features, we also rely on Gabor filters for regional image feature extraction. Gabor filters may be used to measure perceptual surface texture in an image [1]. Specifically, Gabor filters respond to regular patterns in a given orientation on a given scale and frequency. A 2D Gabor filter is given by:

$$\widetilde{G}(x,y) = G_\sigma(x,y) \exp\left\{2\pi i \begin{pmatrix}\Omega_{x_0}\\ \Omega_{y_0}\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}\right\}, \quad i=\sqrt{-1}, \qquad (2)$$

where $G_\sigma(x,y)$ is a Gaussian with a scale σ, $\sqrt{\Omega_{x_0}^2 + \Omega_{y_0}^2}$ is the radial center frequency and $\tan^{-1}(\frac{\Omega_{y_0}}{\Omega_{x_0}})$ the orientation. Note that a zero-frequency Gabor filter reduces to a Gaussian filter.

2.3.3 Keypoint Image Feature Extraction

Inspired by the work of Zhang [27], we also compute invariant descriptors based on interest regions. In an evaluation of interest region detectors, Mikolajczyk et al [14] found that the Harris-Affine detector performs best. However, Zhang obtains best results using the Harris-Laplace interest region detector, noting that affine invariance can often be unstable in the presence of large affine or perspective distortions.

The Harris-Laplace interest region detector [12] uses a Harris corner detector on an image at multiple smoothing scales to detect keypoints. We compute the Laplacian at scales near the scale at which the keypoint was detected. The scale at which the Laplacian is at a local maximum is selected as the scale of the keypoint. The point is rejected if there is no local maximum of the Laplacian. Detected scale

and keypoint together form a circular interest region, which can be detected under rotation and scale changes.

The SIFT descriptor [11] is consistently among the best performing interest region descriptors [13, 27]. SIFT describes the local shape of the interest region using edge histograms. To make the descriptor invariant, while retaining some positional information, the interest region is divided into a 4x4 grid and every sector has its own edge direction histogram (8 bins). The grid is aligned with the dominant direction of the edges in the interest region to make the descriptor rotation invariant.

2.4 Visual-Only Challenge Results

We performed several experiments against the MediaMill Challenge using the various feature vectors in combination with SVM, logistic regression and Fisher's linear discriminant. In addition to using the global, regional, and keypoint features separately, we also explored their combined influence on concept detection performance using vector concatenation.

The Challenge baseline is the SVM with regional Wiccest features, yielding a mean average precision (MAP) of 0.250 on the 39 TRECVID concepts [21]. Our best overall results are obtained with an SVM and regional feature combination using episode constrained cross-validation and inclusion of the γ parameter. Improving upon the Challenge by 41%. Combining features with an SVM yields better performance than using logistic regression or Fisher's linear discriminant. However, these two classifiers allow for quick classification of relatively long feature vectors. Sometimes even outperforming the best SVM detector for a concept. The Fisher linear discriminant is especially effective in classification tasks that involve long feature vectors. When we select the feature and classifier combination that yields the best performance per concept we may obtain an increase over the Challenge of as much as 48% for the 39 TRECVID concepts. For the complete lexicon of 101 concepts from the Challenge the increase is more than 68% (data not shown).

2.5 Submitted Concept Detection Results

All our experiments were performed on the MediaMill Challenge, including parameter optimization and best-of selection. Since the Challenge is based on TRECVID 2005 training data only, we extended the annotations for our final submission with more positive examples from the TRECVID 2005 test set. These were obtained by manual inspection of last years result. We added the positive feature vectors at model construction time, they were not used for parameter optimization. An overview of our submitted concept detection results is depicted in Fig. 3. We will now highlight the details of each submitted run.

2.5.1 Run 'strange': Best Visual-Only

Concept detection that relies on a single feature/classifier combination seldom leads to excellent performance. For some concepts, however, performance is reasonable, e.g., *meeting*, *desert*, *mountain*, *us flag*, *people marching*, *maps*, and *charts*. Our other runs more or less extend on this run to see how performance is influenced by: using concepts in context, adding text, comparison against a keypoint-only run, using cluster-based similarity, and late fusion of several visual-only analysis methods.

2.5.2 Run 'charm': Visual Context Analysis

The context analysis step adds context to our interpretation of the video. Here we combine the best visual-only concept analysis method per concept. The best visual-only run yields a probability for each shot and all 101 concepts detectors in our thesaurus. The probability indicates whether a concept is present.

The results do not show a clear overall advantage of using context for concept detection. For concepts as *mountain*, *corporate leader*, and *military personnel*, context improves upon the best visual-only run. Context aids especially to disambiguate between *maps* and *charts*. For the other concepts the benefit of context is less apparent, but this might be caused by the fact that validation set $\mathcal{B}$ contains less examples than training set $\mathcal{A}$.

2.5.3 Run 'up': Early Fusion

For the 'up' run we performed an early fusion scheme similar to Challenge experiment 3. We combine the feature vectors resulting from visual feature extraction with those obtained from textual feature extraction using vector concatenation.

For the visual features we selected the good performing regional feature combination. To obtain text features, we transformed the ASR text in three ways. The first transformation was pure normalization, eliminating punctuation and capitalization. The second transformation was stemming, using the Porter [17] stemmer to reduce the number of morphological variants of words. The third transformation of the text was character 4-grams, using consecutive sequences of 4 characters for search, to catch 'sounds-like' errors made by the speech transcriber. We used relevance feedback to select the most descriptive *n* terms for each concept. We did this by calculating Rocchio's weight for all of the terms of the ASR text of the positive concept examples, as described in [19].

Early fusion performs reasonably well for *sports*, *waterscape*, *police/security*, *airplane*, and *charts*. Apparently,

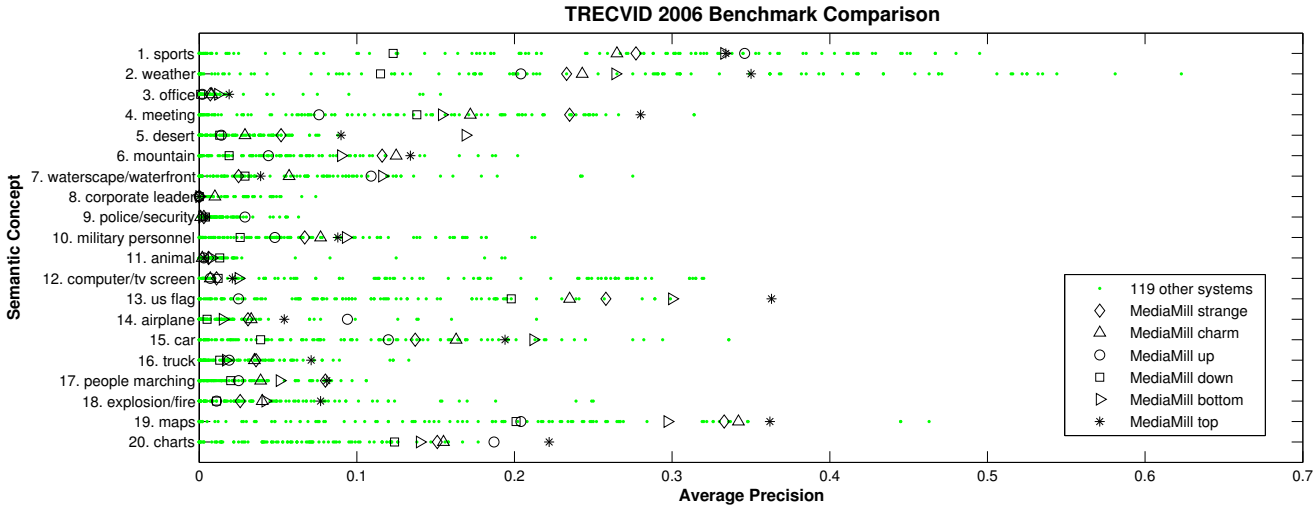

Figure 3: Comparison of MediaMill video indexing experiments with present-day indexing systems in the TRECVID 2006 benchmark.

the text complements the visual features for these concepts. However, for the other concepts addition of text has a negative influence on concept detection performance. In such cases as *meeting*, *desert*, *mountain*, *us flag*, and *maps* resulting in poor performance when compared to our best results for these concepts. Early fusion suffers from textual features based on poor quality (machine translated) ASR.

2.5.4 Run 'down': Late Fusion of Keypoint Detectors

We have selected 5 combinations of detectors and descriptors based on experiments with the MediaMill Challenge:

- Harris-Laplace, SIFT
- Harris-Laplace, Hue *and* SIFT
- Boosted ColorHarris-Laplace, SIFT
- Boosted ColorHarris-Laplace, Hue
- Harris-Laplace *and* Boosted ColorHarris-Laplace, SIFT

For each of the five combinations of interest region detectors and descriptors we have applied SVM, yielding five ranked lists of shots. Shots in the list have a likelihood (provided by the SVM) and naturally the shots with the highest likelihood are ranked at the top. For late fusion of such ranked lists several methods exist, e.g., min, max, sum, median, and product [5]. An extension of product fusion that is capable to handle missing data is the geometric mean. We found after several experiments on Challenge data that this geometric mean outperforms the other fusion methods. Hence, we combine the various lists using the geometric mean.

Visual inspection of results shows that there are many topics where many top ranked results do not look like the target concept at all (from a human perspective, at least). However, there is a pattern in those results: they all tend to have many smooth areas, be relatively blurred and/or lack saturated colors. These are all conditions in which an interest region detector will detect few interest regions. Looking at the results of our Harris-Laplace interest region detector we can see that there are many keyframes with few interest regions in this run. For the top 100 shots of the runs of all concepts evaluated this year, 30% have 10 interest regions or less in run 5. In all other runs it does not exceed 10%. One might be tempted to remove shots with few interest regions because they introduce many incorrect results, but this can have side-effects. For the first 100 shots of the concept animal, 10 shots have been evaluated as correct. However, five of these have less than 10 interest regions. Removing these shots would cause a serious decrease in performance for this concept. We are currently investigating how to handle keyframes with few detected interest points.

2.5.5 Run 'bottom': Proto-Concept Clustering

This run constructs a dictionary of proto-concepts for the Weibull and Gabor features in a data-driven approach. This data-driven approach was developed in parallel to the other experiments. Hence, this run is not incorporated in the fusion, best-visual or context runs. Nevertheless, the data-driven approach outperforms the other MediaMill runs for 6 out of 20 concepts. Moreover, the concept *desert* yields the best result over all other systems. Hence, a data-driven

approach for finding a dictionary of proto-concepts complements the other runs and even yields first-rate performance for some concepts.

2.5.6 Run 'top': Late Fusion of Visual-Only Analysis

This run is a late fusion of all our experiments based on visual features. For the 39 TRECVID concepts all experiments from [21] and the keypoint feature run ('down') are included. However, fusing *all* experiments did not yield good results on Challenge data. Instead, we choose to use a variable number of experiments per concept. The combination always includes the keypoint feature run as an experiment. The combination method adds further experiments on a per-concept basis. Experiments are added in order of decreasing performance. We consider combinations of up to 10 experiments. Per concept we select the number of experiments that yields the best average precision performance on Challenge validation set $\mathcal{B}$. The fusion of the different experiments is again performed using the geometric mean.

The fusion of visual-only analysis results is our best overall run. Moreover, we obtain the highest performance for pure visual concepts *flag us* and *charts*. We also perform well for concepts *meeting*, *desert*, and *maps*. For concepts with relatively few learning examples, e.g., *corporate leader* and *police/security*, classification remains hard. Relative to other concept detection methods we perform poor for *computer/tv screen*. This is caused, however, by the fact that we do not consider screens that appear in a news studio setting as valid examples. Since detection here boils down to detecting the studio or news anchor. It is interesting to note that fusion always outperforms the best single visual-only analysis approach, except for *animal* where both scores are close to zero. The 'bottom' run was not included in the fusion, inclusion of this run in the fusion will further improve concept classification performance.

2.6 Scaling-up to 491 Concept Detectors

To scale our lexicon of concept detectors further we adopt a graceful degradation approach. For the remaining 62 MediaMill concepts, the keypoint features from the 'down' run and the SVM gamma experiment are not available. We determine the best combination of experiments for these concepts from the remaining experiments; again up to 10 experiments are allowed in a combination. For the LSCOM concepts [15] none of the SVM experiments are available, leading to a further reduction in the number of experiments, i.e. only those performed by logistic regression and Fisher's linear discriminant. Because parameter optimization of the SVM is expensive – even when supercomputers are used – performing a complete analysis for all concepts was not feasible. While the performance might not be optimal, the detectors may still be useful for semantic video retrieval.

3 Semantic Video Retrieval

Our TRECVID 2006 search task efforts have concentrated on interactive retrieval using the lexicon of 491 learned concept detectors. Query by concept yields a ranking of the data, a convenient way of browsing the result is our CrossBrowser [23] which allows to use both the rank and temporal context of a shot. There are, however, many other relevant directions which can be explored e.g. different semantic threads through the data or shots visually similar to the current shot. This year we therefore developed the RotorBrowser which allows the user to browse along eight directions. We depict both browsers in Fig. 4.

We submitted two runs for interactive search with two expert users. One user performed the interactive search by using the MediaMill search engine with the CrossBrowser. Another user exploited the MediaMill system in combination with the RotorBrowser. Results in Fig. 5 indicate that for most search topics, users of the MediaMill system score above average. Furthermore, users of our approach obtain a top-3 average precision result for 14 out of 24 topics. Best performance is obtained for 6 topics. Among all interactive video retrieval systems the CrossBrowser ranked 2nd and the RotorBrowser 6th.

The CrossBrowser is especially successful when a search topic can be addressed with a single concept detector from the lexicon. Finding a helicopter in flight, for example, is relatively easy when a reasonably accurate *helicopter* detector is available. The CrossBrowser then allows for quick scanning and selection of relevant results. When search topics contain combinations of several reliable concept detectors, e.g. *people*, *suits*, *flag* (Topic: 17), results are not optimal. This indicates that much is to be expected from a more intelligent combination of query results. Overall we can say that the RotorBrowser is able to find similar results as the CrossBrowser with less user interaction. However, more tuning is required to make it visualize relevant threads only. A more in-depth study using a larger (novice) user base is currently underway to determine the possible benefit of having multiple dimensions in browsing.

4 Conclusion

In this paper we have presented a state of the art video search engine. It relies on powerful keypoint-local, regional and global visual features, machine learned concepts, and interaction. The software is consolidated in a *C++* library of functions and tasks suited for beta use. We will demo semantic video search with the MediaMill system at the conference.

Although the current interaction mechanism is an asset, it can still be improved to be more effective. The same holds for the shot segmentation. The current system lacks

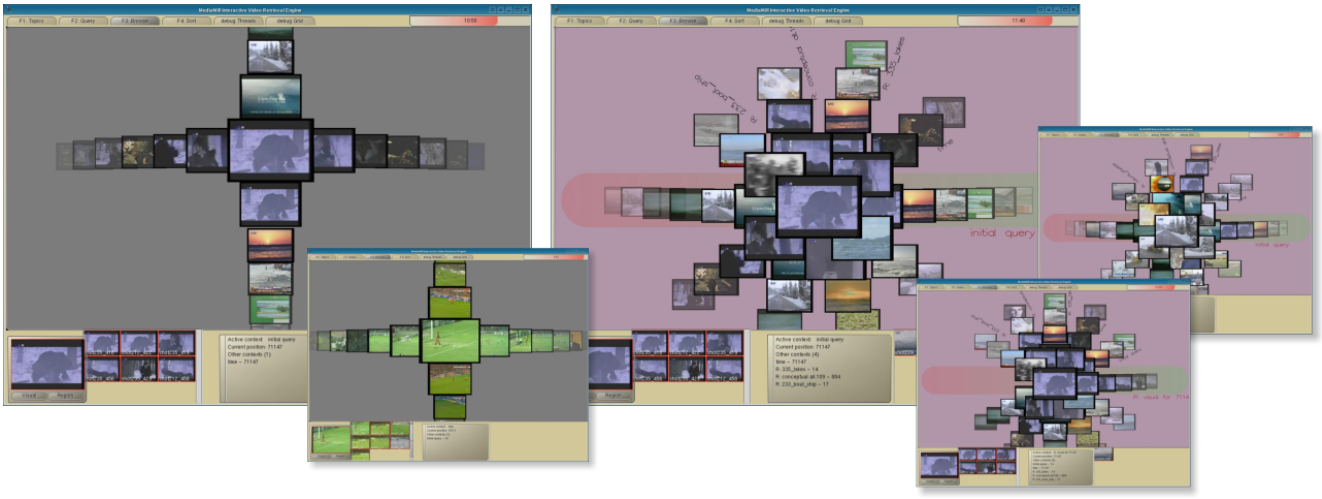

Figure 4: Screenshots of the CrossBrowser (left) and the RotorBrowser.

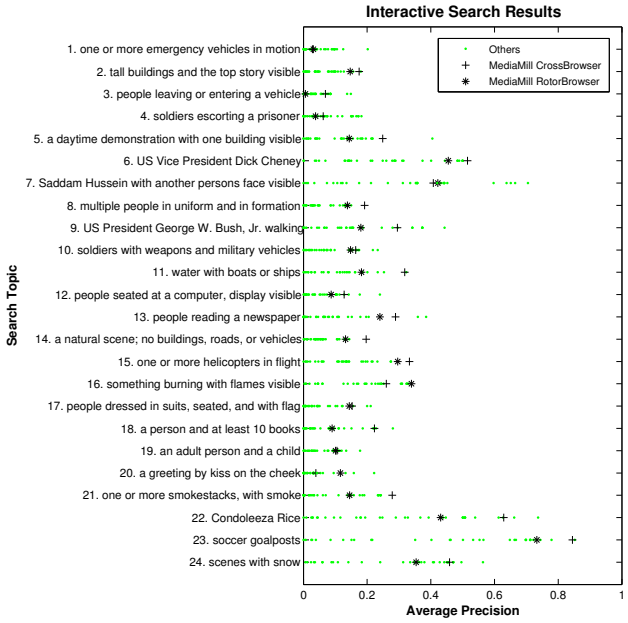

Figure 5: Comparison of interactive video search results for 24 topics performed by 36 users of present-day video retrieval systems. MediaMill results are indicated with special markers.

advanced machine learning, audio events, features based on tracking, speech and text analysis, and ontology-driven query expansion. These elements are part of ongoing research in the national MUNCH and European VIDI-Video project. We are grateful these projects are based on a collaboration with many, well-respected scientists to achieve the scale needed to advance video search. The system we present today performed among the best three in the international TRECVID competition for the third year in a row.

Acknowledgments

This research is sponsored by the BSIK MultimediaN project, the NWO MuNCH project, and the EU 6th Framework project VIDI-Video.

References

[1] A. C. Bovik, M. Clark, and W. S. Geisler. Multichannel texture analysis using localized spatial filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(1):55–73, 1990.

[2] 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/`.

[3] R. P. W. Duin et al. PRTools version 4.0: A matlab toolbox for pattern recognition, 2006. `http://www.prtools.org/`.

[4] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[5] E. Fox and J. Shaw. Combination of multiple searches. In *TREC-2*, pages 243–252, 1994.

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

[7] 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 *Int'l Workshop on Semantic Learning Applications in Multimedia, in conjunction with CVPR'06*, New York, USA, June 2006.

[8] J. C. van Gemert, C. G. M. Snoek, C. J. Veenman, and A. W. M. Smeulders. The influence of cross-validation on video classification performance. In *Proceedings of the*

ACM International Conference on Multimedia, pages 695–698, Santa Barbara, USA, October 2006.

[9] J.-M. Geusebroek and A. W. M. Smeulders. A six-stimulus theory for stochastic texture. *International Journal of Computer Vision*, 62(1/2):7–16, 2005.

[10] J.-M. Geusebroek. Compact object descriptors from local colour invariant histograms. In *British Machine Vision Conference*, Edinburgh, UK, September 2006.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[12] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[13] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005.

[14] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65(1-2):43–72, 2005.

[15] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, 2006.

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

[17] M. F. Porter. An algorithm for suffix stripping. In *Readings in Information Retrieval*, pages 313–316. Morgan Kaufmann, San Francisco, CA, 1997.

[18] G. Quénot, D. Moraru, L. Besacier, and P. Mulhem. CLIPS at TREC-11: Experiments in video retrieval. In E. Voorhees and L. Buckland, editors, *Proceedings of the 11th Text REtrieval Conference*, volume 500-251 of *NIST Special Publication*, Gaithersburg, USA, 2002.

[19] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. In *Readings in Information Retrieval*, pages 355–364. Morgan Kaufmann, San Francisco, CA, 1997.

[20] C. G. M. Snoek, J. C. van Gemert, J.-M. Geusebroek, B. Huurnink, D. C. Koelma, G. P. Nguyen, O. de Rooij, F. J. Seinstra, A. W. M. Smeulders, C. J. Veenman, and M. Worring. The MediaMill TRECVID 2005 semantic video search engine. In *Proceedings of the 3rd TRECVID Workshop*, Gaithersburg, USA, November 2005.

[21] 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 *Proceedings of the 4th TRECVID Workshop*, Gaithersburg, USA, November 2006.

[22] 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 Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1678–1689, October 2006.

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

[24] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM International Conference on Multimedia*, pages 421–430, Santa Barbara, USA, October 2006.

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

[26] H. Wactlar, M. Christel, Y. Gong, and A. Hauptmann. Lessons learned from building a terabyte digital video library. *IEEE Computer*, 32(2):66–73, 1999.

[27] 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.