THE SEMANTIC PATHFINDER FOR GENERIC NEWS VIDEO INDEXING

C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma, F.J. Seinstra, and A.W.M. Smeulders

Intelligent Systems Lab Amsterdam, University of Amsterdam,
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
{cgmsnoek, worring, mark, koelma, fjseins, smeulders}@science.uva.nl

ABSTRACT

This paper presents the semantic pathfinder architecture for generic indexing of video archives. The pathfinder automatically extracts semantic concepts from video based on the exploration of different paths through three consecutive analysis steps, closely linked to the video production process, namely: content analysis, style analysis, and context analysis. The virtue of the semantic pathfinder is its learned ability to find a best path of analysis steps on a per-concept basis. To show the generality of this indexing approach we develop detectors for a lexicon of 32 concepts and we evaluate the semantic pathfinder against the 2004 NIST TRECVID video retrieval benchmark, using a news archive of 64 hours. Top ranking performance indicates the merit of the semantic pathfinder.

1. INTRODUCTION

Query-by-keyword forms the foundation for access to text repositories. Elaborating on the success of text-based search engines, query-by-keyword is also the paradigm of choice in multimedia retrieval scenarios. For multimedia archives it is hard to achieve effective access, however, when based on keywords that appear in the text only. Video archives require semantic access where all modalities can contribute to the concept. In literature a varied gamut of specific multimedia keyword, or *concept*, detectors haven been proposed; where concepts like *tigers* and *sunsets* are prototypical examples. Although specific methods have aided in achieving progress, this road is a dead end given the plethora of concepts which are needed for effective access. It is simply impossible to design a tailor-made solution for each concept.

In this paper, we propose a generic approach for concept indexing, we call the *semantic pathfinder*. The design principle of the semantic pathfinder is derived from the observation that video indexing can be regarded as the inversion of video production, covering notions of content, style [10], and context [6]. The semantic pathfinder exploits analysis steps at increasing levels of abstraction, corresponding to well-known facts from the study of films and television production [2]. While doing so, it combines the most successful methods for semantic video indexing [1, 4, 6, 10, 12] into an integrated architecture. In contrast to these methods, however, we do not trust blindly on one single technique for generic video indexing. One would expect that some concepts, like *vegetation*, have their emphasis on content where the style (of the camera work that is) and context (of concepts like *graphics*) do not add much. In contrast, more complex events, like *people walking*, profit from incremental adaptation of the analysis to the intention of the author. Hence, we advocate that the best indexing approach is concept-dependent. The virtue of the semantic pathfinder is its learned ability to find the best path of analysis steps on a per-concept basis. To demonstrate the effectiveness of the semantic pathfinder, the semantic indexing experiments are evaluated within the 2004 NIST TRECVID video retrieval benchmark [8].

The organization of this paper is as follows. First, we introduce the semantic pathfinder. The experimental setup is explained in Section 3. We present results in Section 4.

2. SEMANTIC PATHFINDER

The essence of produced video, like broadcast news or feature films, is that an author creates the final program. It is more than just the content. Before creation, the author starts with a semantic idea: an interplay of concepts and context. To stress the semantics of the message, guiding the audience in its interpretation, the author combines various stylish production facets, such as camera framing and synchronization of voice-overs with visual content. The video aims at an effective semantic communication. Hence, the core of semantic indexing is to reverse this authoring process. We follow this path to arrive at a system architecture for semantic indexing in video. Before we elaborate on the video indexing architecture, we first define a lexicon Λ_S of 32 semantic concepts, see Fig. 1. We aim to detect all 32 concepts with the proposed system architecture.

The semantic pathfinder is composed of three analysis steps. It follows the reverse authoring process. Each analysis step in the path detects semantic concepts, but each from a different authoring perspective. In addition, one can exploit the output of an analysis step in the path as the input for the

An extended version of this paper will appear in IEEE Trans. PAMI. This research is sponsored by TNO and the BSIK MultimediaN project.

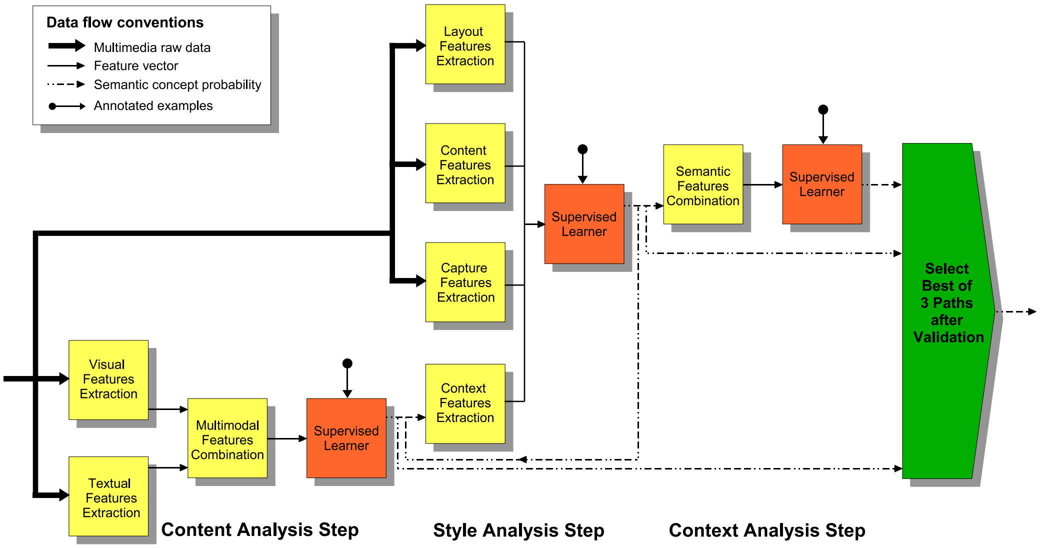
Data flow conventions
Multimedia raw data
Feature vector
Semantic concept probability
Annotated examples
Layout Features Extraction
Content Features Extraction
Capture Features Extraction
Context Features Extraction
Visual Features Extraction
Textual Features Extraction
Multimodal Features Combination
Supervised Learner
Supervised Learner
Semantic Features Combination
Supervised Learner
Select Best of 3 Paths after Validation
Content Analysis Step
Style Analysis Step
Context Analysis Step

Fig. 2. The semantic pathfinder for generic video indexing.

next one. The analysis steps in the semantic pathfinder exploit a common architecture with a standardized input-output model to allow for semantic integration. An overview of the semantic pathfinder is given in Fig. 2. The various components are now explained in more detail.

2.1. Supervised Learning Module

We perceive semantic indexing in video as a pattern recognition problem. Given pattern x, describing the video from one of the authoring perspectives at the granularity of a shot i, the aim is to obtain a confidence measure, which indicates whether semantic concept ω is present in shot i. Each analysis step in the semantic pathfinder extracts x_i from the data, and exploits a Support Vector Machine (SVM) [11] to learn $p(\omega|x_i)$ for all ω in the semantic lexicon Λ_S from labeled examples. The SVM margin is converted to a probability using a sigmoid function. To obtain good settings, $\vec{q}^*$ for the SVM, we perform parameter search on a large number of SVM parameter combinations. The result of the parameter search over $\vec{q}$ results in the model $p(\omega|x_i, \vec{q}^*)$ specific for ω. We split the training data, including provided labeled examples, a priori into a non-overlapping training set (85%) and validation set (15%) to prevent overfitting of classifiers.

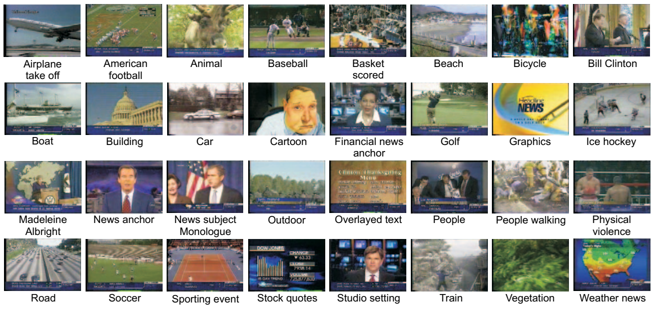

Fig. 1. Instances of the 32 concepts in the lexicon, as detected with the semantic pathfinder.

2.2. Content Analysis Step

The semantic pathfinder starts in the *content analysis step*. In this analysis step, we follow a data-driven perspective on indexing semantics. Here we summarize the approach, for details we refer to [9]. For visual feature extraction we first extract a number of invariant visual features per pixel. Based on these features the procedure labels each pixel in an image with one of 18 low-level visual concepts, like *concrete*, *sand*, *sky*, and so on. This pixel-wise classification results in a labeled segmentation of a key frame in terms of regional visual concepts. The percentage of pixels associated to each of the 18 visual concepts is used as a visual content vector $\vec{v}_i$. In the textual modality, we learn the relation between uttered speech [3] and concepts. We connect words to shots and derive a lexicon of uttered words that co-occur with ω, yielding Λ_ω. For feature extraction we compare the text associated with each shot with Λ_ω. This comparison yields a text vector $\vec{t}_i$ for shot i, which contains the histogram of the words in association with ω. We concatenate $\vec{v}_i$ with $\vec{t}_i$. After feature normalization, we obtain fusion vector $\vec{f}_i$. Then $\vec{f}_i$ serves as the input for the supervised learning module, which learns the semantic concept for the content analysis step.

2.3. Style Analysis Step

In the *style analysis step* we conceive of a video from the production perspective. Based on the four roles involved in the

Table 1. Test set precision at 100 after the three steps, for a lexicon of 32 concepts. The optimal selected path by the semantic pathfinder is given in bold.

Semantic Concept	*Content Analysis*	*Style Analysis*	*Context Analysis*	*Semantic Concept*	*Content Analysis*	*Style Analysis*	*Context Analysis*
News subject monologue	0.55	**1.00**	1.00	Baseball	**0.54**	0.43	0.47
Weather news	**1.00**	1.00	1.00	Building	**0.53**	0.46	0.43
News anchor	0.98	0.98	**0.99**	Road	0.43	0.53	**0.51**
Overlayed text	0.84	**0.99**	0.93	American football	**0.46**	0.18	0.17
Sporting event	0.77	**0.98**	0.93	Boat	0.42	0.38	**0.37**
Studio setting	0.95	0.96	**0.98**	Physical violence	0.17	0.25	**0.31**
Graphics	0.92	0.90	**0.91**	Basket scored	0.24	0.21	**0.30**
People	0.73	0.78	**0.91**	Animal	0.37	0.26	**0.26**
Outdoor	0.62	0.83	**0.90**	Bill Clinton	**0.26**	0.35	0.37
Stock quotes	**0.89**	0.77	0.77	Golf	**0.24**	0.19	0.06
People walking	0.65	0.72	**0.83**	Beach	0.13	0.12	**0.12**
Car	0.63	0.81	**0.75**	Madeleine Albright	**0.12**	0.05	0.04
Cartoon	0.71	0.69	**0.75**	Airplane take off	0.10	0.08	**0.08**
Vegetation	**0.72**	0.64	0.70	Bicycle	0.09	**0.08**	0.07
Ice hockey	**0.71**	0.68	0.60	Train	**0.07**	0.07	0.03
Financial news anchor	0.40	**0.70**	0.71	Soccer	**0.01**	0.01	0.00

video production process [9, 10], this step analyzes a video by four related style detectors. Layout detectors analyze the role of the editor. Content detectors analyze the role of production design. Capture detectors analyze the role of the production recording unit. Finally, context detectors analyze the role of the preproduction team.

Extensive implementation details of the various detectors are in [9, 10]. We restrict ourselves here to an enumeration. The set of layout features is given by: $\mathcal{L} = \{$*shot length, overlayed text, silence, voice-over*$\}$. The set of content features is given by: $\mathcal{C} =\{$*faces, face location, cars, object motion, frequent speaker, overlayed text length, video text named entity, voice named entity*$\}$. The set of capture features is given by: $\mathcal{T} = \{$*camera distance, camera work, camera motion*$\}$. The basic set of context features is given by: $\mathcal{S} = \{$*news reporter, content analysis step* $\omega_1\}$. Where ω_1 indicates the concept from Λ_S with the best average precision performance on the validation set after the context analysis step. The concatenation of $\{\mathcal{L}, \mathcal{C}, \mathcal{T}, \mathcal{S}\}$ for shot i yields style vector $\vec{s}_i$. This vector forms the input for the supervised learning module, which trains a style model for each concept in Λ_S in an iterative fashion. In addition, it forms the input for the next analysis step in our semantic pathfinder.

2.4. Context Analysis Step

The *context analysis step* adds context to our interpretation of the video. Here our aim is the reconstruction of the author's intent by considering detected concepts in context [6]. We use the 32 scores from the style analysis as semantic features. We fuse them into context vector, $\vec{c}_i$. From $\vec{c}_i$ we learn relations between concepts automatically. To that end, $\vec{c}_i$ serves as the input for the supervised learning module, which associates a contextual probability $p(\omega|\vec{c}_i, \vec{q}^*)$ to a shot i for all ω in Λ_S.

The output of the context analysis step is also the output of the entire semantic pathfinder on news video. On the way we have included in the pathfinder, the results of the analysis on raw data, facts derived from production by the use of style features, and an intentional perspective of the author's objective by using concepts in context. For each concept we obtain a probability based on content, style, and context. The semantic pathfinder selects from the three possibilities the one that maximizes average precision based on validation set performance.

3. EXPERIMENTAL SETUP

To demonstrate the effectiveness of the semantic pathfinder, we have participated in the semantic concept detection task of the 2004 NIST TRECVID video retrieval benchmark [8]; the *de facto* standard to evaluate performance of video indexing and retrieval research. The video archive of the 2004 TRECVID benchmark is composed of 184 hours of US News from 1998 and is recorded in MPEG-1 format. The training data contains approximately 120 hours, the test data contains the remaining 64 hours. Together with the video archive, CLIPS-IMAG [7] provided a camera shot segmentation. For the annotations we rely in part on the provided ground truth in TRECVID 2003 [5]. We remove the many errors from this annotation effort. It is extended manually to arrive at an incomplete, but reliable ground truth[1] for all concepts in lexicon Λ_S. To determine the accuracy of concept detection we use *precision at 100*, following the standard in TRECVID evaluations [8]. The TRECVID 2004 procedure prescribes that only 10 pre-defined concepts are evaluated by NIST. For these 10 concepts we report the official benchmark results using *average precision* [8].

4. RESULTS

We evaluated detection results for all 32 concepts in each analysis step. The *precision at 100* is reported in Table 1. We observe from the results that the learned best path (printed in bold) indeed varies over the concepts. The virtue of the

[1] Available: `http://www.science.uva.nl/~cgmsnoek/tv/`.

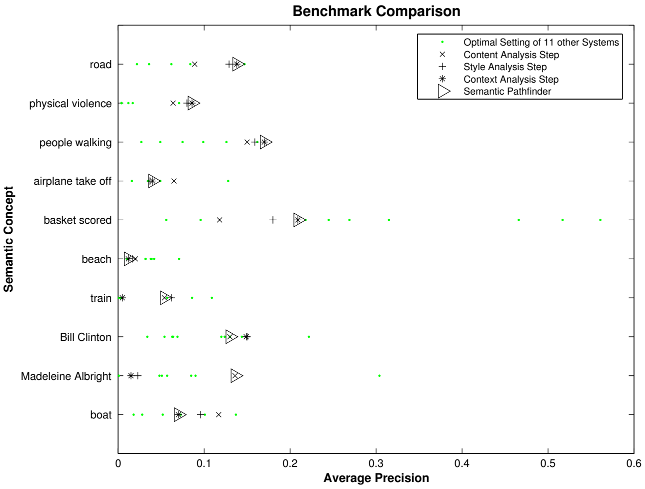

Fig. 3. Comparison of semantic pathfinder results with 11 other indexing systems in the TRECVID 2004 benchmark.

semantic pathfinder is demonstrated by the fact that concepts are indeed divided by the analysis step after which they achieve best performance. For 12 concepts, the learning phase indicates it is best to concentrate on content only. For 5 concepts, the semantic pathfinder demonstrates that a two-step path is best (where in 15 cases addition of style features has a marginal positive or negative effect). For 15 concepts, the context analysis step obtains a better result, where in 5 cases this leads to a substantial increase. Thus, some concepts are just content, style does not affect them. In such cases as *American football* there is style-wise too much confusion with other sports to add new value in the path. Style does help when the concepts are semantically rich: e.g. *news subject monologue* and *sporting event*. For complex concepts, analysis based on content and style is not enough. They require the use of context. The context analysis step is especially good in detecting named events, like *people walking*, *physical violence*, and *basket scored*.

We performed an additional experiment within the TRECVID benchmark to compare the effectiveness of the semantic pathfinder for detection of concepts to 11 present-day video indexing systems. We select from each participant the system tuning with the best performance for a concept out of a maximum of 10 tunings. Results are visualized in Fig. 3 for each concept. Relative to other video indexing systems the semantic pathfinder performs the best for two concepts, i.e. *people walking* and *physical violence*, and second for five concepts. For two concepts we perform moderate, i.e. *basket scored* and *beach*. Here the best approaches are based on specialized concept detection methods that exploit domain knowledge. The big disadvantage of these methods is that they are specifically designed and implemented for one concept. They do not scale to other concepts. The benchmark results show that the semantic pathfinder allows for generic indexing with state-of-the-art performance.

5. CONCLUSION

We propose the semantic pathfinder, a generic approach for video indexing. It is based on the observation that produced video is the result of an authoring process. Experiments with a lexicon of 32 semantic concepts demonstrate that the semantic pathfinder allows for generic video indexing, while confirming the value of the authoring metaphor in indexing. In addition, the results over the various analysis steps indicate that a technique taxonomy exists for solving concept detection tasks; depending on whether content, style, or context is most suited for indexing. Finally, the semantic pathfinder is successfully evaluated within the 2004 TRECVID benchmark. With one and the same set of system parameters two concepts came out best against 11 other present-day systems. For five concepts our system scored second best. Just two performed poorly in this comparison. The results show that the semantic pathfinder allows for state-of-the-art performance without the need of implementing specialized detectors. We consider this the best indication of the validity of the approach.

6. REFERENCES

[1] A. Amir et al. IBM research TRECVID-2003 video retrieval system. In *Proc. TRECVID Workshop*, USA, 2003.

[2] D. Bordwell and K. Thompson. *Film Art: An Introduction*. McGraw-Hill, New York, USA, 5th edition, 1997.

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

[4] A. Hauptmann et al. Informedia at TRECVID 2003: Analyzing and searching broadcast news video. In *TRECVID Workshop*, Gaithersburg, USA, 2003.

[5] C.-Y. Lin, B. Tseng, and J. Smith. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In *TRECVID Workshop*, USA, 2003.

[6] M. Naphade and T. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Transactions on Multimedia*, 3(1):141–151, 2001.

[7] G. Quénot et al. CLIPS at TREC-11: Experiments in video retrieval. In *Text REtrieval Conf.*, Gaithersburg, USA, 2002.

[8] A. Smeaton, P. Over, and W. Kraaij. TRECVID: Evaluating the effectiveness of information retrieval tasks on digital video. In *ACM Multimedia*, NY, USA, 2004.

[9] C. Snoek. *The Authoring Metaphor to Machine Understanding of Multimedia*. PhD thesis, Univ. van Amsterdam, 2005.

[10] C. Snoek, M. Worring, and A. Hauptmann. Learning rich semantics from news video archives by style analysis. *ACM TOMCCAP*, 2(2), 2006. In press.

[11] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, NY, USA, 2th edition, 2000.

[12] 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.