Arnold W.M. Smeulders, Jan van Gemert, Jan-Mark Geusebroek, Cees Snoek, and Marcel Worring

Browsing for the national Dutch video archive

Abstract— **Pictures have always been a prime carrier of Dutch culture. But pictures take a new form. We live in times of broad- and narrowcasting through Internet, of passive and active viewers, of direct or delayed broadcast, and of digital pictures being delivered in the museum or at home. At the same time, the picture and television archives turn digital. Archives are going to be swamped with information requests unless they swiftly adapt to partially automatic annotation and digital retrieval. Our aim is to provide faster and more complete access to picture archives by digital analysis. Our approach consists of a multi-media analysis of features of pictures in tandem with the language that describes those pictures, under the guidance of a visual ontology. The general scientific paradigm we address is the detection of directly observables fused into semantic features learned from large repositories of digital video. We use invariant, natural-image statistics-based contextual feature sets for capturing the concepts of images and integrate that as early as possible with text. The system consists of a large for science yet small for practice set of visual concepts permitting the retrieval of semantically formulated queries. We will demonstrate a PC-based, off-line trained state of the art system for browsing broadcast news-archives.**

I.INTRODUCTION

Pictorial information has always been the preferred carrier of Dutch and Flemish culture. Where some cultures drive on music, cooking, martial arts, story telling, poetry, or dance, others have a preference for visual expression. There runs a thread of painters through the history of Dutch culture as illustrated by the sequence of painters from Jeroen Bosch, Van Eijck, Pieter Brueghel, Rembrandt and his disciples, Frans Hals, Pieter Vermeer, Steen, Potter, the Dutch landscape painters, the Hague school, Van Gogh, Mondriaan up to William de Kooning in the current day. Other forms of pictorial communication such as television commercials and industrial design are similarly important.

In recognition, museums for visual arts are in high esteem in the Netherlands. Only shortly after the start of television, a national archive for television was formed: Beeld en Geluid [translated into: Vision and Sound]. It holds 700,000 hours of film, video and audio, one of the largest in its kind in the world.

With the advent of digital data, the institute wishes to convert to an all digital handling of all broadcasts in the Netherlands, some 10,000 hours of material each year. This poses a considerable burden on the capability of the staff to annotate all of this for future access, be it professional or for the general public.

At the same time, we live in exciting times of the advent of broad- and narrowcasting through the Internet, of viewers passively and actively engaged in the broadcast, of direct or delayed broadcast, of the advent of hundreds of television channels, and of digital pictures being delivered anywhere, nowhere and at home. Picture and television archives turn digital.

In these demanding times, the archives will be swamped with requests to reuse previous broadcast material. Unless the archive adapts swiftly to semi-automatic annotation and digital retrieval it will be lost in cyberspace.

In the MuNCH-project, we aim to bring relieve to the archive by focusing on combining visual with speech converted to textual information under the guidance of explicit knowledge for the purpose of providing automated indices and interactive access by content-based image and text retrieval [Smeulders 2000].

The general scientific question of the project reads: *Can directly visual observables and early integration of text and image features solve the main issues in providing access to digital video archives?* Is it possible to learn semantic concepts from mixed media video to a sufficient degree to provide broad access? These questions translate into: Can weak concept classifiers be integrated in retrieval of semantically formulated queries? How can retrieval methods for language best be expanded into an ontology for visual observables? What kind of visual properties should be added to the (limited) set of concepts with a visual component, such that information can be employed for either weak recognizers and/or text-image fusion techniques?

II.CONRIBUTION TO LIBRARY SCIENCE

The digital production and the sheer amount of broadcast data poses formidable challenges that cannot be handled without the support of automated analysis tools. While recognizing that the current practice cannot be left unchanged under automated analysis [Smeulders 2004], we depart from the view that *new technology is always first accepted in the old idiom*. Beeld & Geluid has analyzed current queries. We use that as our leading principle to build a proper workflow around the digital archive, or for that matter, an effective digital system around the archivist.

In digital access to a digital library, the key point is to *differentiate in search patterns*. What was perfect for manual search is no longer the case for automated access. One should be aiming well outside key-word search. In *query from a controlled vocabulary,* specification of the query is from a given set already pre-computed in the archive. With automatically indexed documents, multidimensional and visual projection helps to provide

The authors are with the Intelligent Systems Lab Amsterdam. Sponsored by the MuNCH-CATCHproject & the MultimediaN project. Correspondence to smeulders@science.uva.nl.

an overview of the data being accessed. In *query by keywords or descriptors* the user has the possibility to query on the content of the archive directly. With uncertain annotations, a gradual shift is anticipated to *query by full text or full visual examples.* Keywords or descriptors entered by the user are very focused, where a machine does not grasp the context by itself, nor does it have experience unless programmed, nor does it have a good sense of the user's search aim either. Therefore, search algorithms profit from a full text retrieval with images included and iterative interaction. Indeed, in practice the user will engage in an interactive session using *relevance feedback, visual presentation* and *question-answer* sessions.

III. Related Work

There are relatively few reports on picture language ontologies and retrieval. One of the early techniques was described in ImageRover [Sclaroff 1998]. This and other early approaches effectively employed a bag-of-words & bag-of-features approach to images and text. A hierarchically ordered set of features is likely to be more effective in finding semantics especially when some fraction of the features is directly observable.
From the more recent literature we mention [Hauptmann 2004] where an effort is proposed to come to a large-scale concept ontology for broadcast video. Rather than identifying a list of frequently mentioned concepts, we aim with first priority on concepts which are either directly bootstrap semantic interpretation or which can be detected by text and pictorial information directly. Our approach supplements the search for a large number of specific sign detectors, equally useful in semantic video retrieval. The papers [Schober 2004] [Maillot 2004] describe tentative picture ontologies employing features, which will be highly dependent on the accidental conditions of the recording. Hence, the retrieval will find the images which pictorial identical. This is a far step from retrieving semantically identical images as we aim for in this research.

IV. Approach

We combine recent advances in computer vision in the area of scene characterization as described by natural image statistics [Geusebroek 2002] with advances in text processing and information retrieval [Jijkoun 2004] under the guidance of a visual ontology [Schreiber 2001].

The fundamental obstacle in automatic annotation is the *semantic gap* between the digital data and their semantic interpretation. The conversion of images to words has intrigued many, from ancient philosophers to modern day neuro-computationalists. The visual-auditory-language processing takes up half of the capacity of a brain, demonstrating that any media processing must be a scientifically and technologically challenging problem.

Text interpretation can be performed at much deeper levels of understanding than before, while maintaining robustness and wide-coverage. This is due to the availability of more training data, more sophisticated machine learning technology, and increased computational resources [Jackson 2002]. Here, we focus exclusively on the contribution of visual features to television annotation as an example of multimedia data streams.

In recent years, significant progress has been made on understanding text by natural language processing and images by computer vision. In machine vision, the use of invariant features has made object recognition much simpler by closing the *sensory gap* between objects and the millions of its slightly different appearances due to differences in illumination, viewing angle and scenes. Retrieving known objects [Geusebroek 2005] has shown excellent progress, while promising results are being achieved in learning object categories [Fergus 2002], to name a few. These results are based on invariant and complete feature sets [Lowe 2005]. The use of invariant feature sets ensures that the accidental conditions introduced in the image at the moment of the recording are removed from the picture. What remains are the object-intrinsic properties desperately needed in the interpretation of the scene.
Apart from being invariant, features also need to complete in the sense that they provide a rich description of the objects and components in the scene. And, the invariant features need to be discriminative to be capable of distinguishing the thousands of visual concepts from one another. The above mentioned feature set [Lowe 2004] is the state of the art in this regard, to be improved by [Geusebroek 2005].

Mining visual and textual data have been carried out in splendid separation, almost exclusively. Knowledge that combines textual and pictorial description is urgent needed to enhance analysis and retrieval of mixed-media to much higher levels. In world-wide, task-based retrieval evaluations such as videoTREC and imageCLEF, an integrated approach combining text and image information at a low level and the use of machine learning techniques is the essential ingredient of the most successful systems [Snoek 2004]. We will pursue this line of research by connecting invariant feature sets with their scene-dependent variance taken out and connecting the remaining information to semantic concept in a visual ontology.

In view of the wealth of information in a video stream, we identify two strategies:

(1) A set of detectors of directly visual-only observables ordered in a visual ontology, Directly observables are descriptors of image patches that permit immediate denomination of that patch even when the small piece is shown without context. Oak and timber in general, foliage, bricks, wool and cloth in general, and stucco are

good examples of directly visual observables. These materials are in our approach to bootstrap the interpretation of the scene. From there we build a visual ontology of semantic features [Hollink 2004ab, Smeulders 2002] including the expectation of what else can be seen in the scene. When sampled with scene-invariant features, they provide a semantic anchor in the image regardless of context.

(2) A set of concept detectors based on by early fusion of visual analysis with text interpretation. In the other line of research, combining visual, audio and text features learns concepts of objects. Experience from previous research [Snoek 2004] has learned that the three modalities to be combined at the lowest level of analysis to be effective.

We exploit the principle of early fusion. As machine learning provides robustness in real-world data analysis, we apply it to build early fused recognition of concepts as well as directly observable detectors. The size of the example set is indicative for the quality of the result. Hence, it is essential to reduce the amount of annotated data by active learning of one-class classifiers. The semantics of a scene will usually be carried in a describing text, unless the text is used in metaphor. Early combination of visual and textual signs as described above as the second line of research may indicate whether a concept is actually in the scene or that the text is just referring to an abstract concept. The text information will enhance the precision and accuracy of mixed-media retrieval.

This results in the system architecture plotted in Fig. 1. In this fashion we have successfully participated in videoTREC 2004, where we achieved top ranking results based on the semantic value chain [Snoek 2004] and achieved the following results, see Fig. 2.

V.DISCUSSION

Understanding words in their ability to summarize the infinitely complex world has intrigued many. By building a visual ontology and by learning early fusion of textual and pictorial information we aim to make a significant contribution in the quest for semantics.

While the fields of content-based image retrieval and of text retrieval are very active [Smeulders 2000], combinations of text and image are still rare. Yet much is to be gained by combinations, both of data from multiple sources and of data from essentially different media (images and text). They can be connected in various ways to give us deeper insights into the nature of objects and processes, the knowledge thereof and the corresponding textual descriptions.

V. REFERENCES

[Schreiber 2001] G. Schreiber, B. Dubbeldam, J. Wielemaker, B. Wielinga. Ontology-based photo annotation *IEEE Intell Systems* 2001.

[Smeulders 2002] A.W.M. Smeulders, L. Hardman, G. Schreiber, J.M. Geusebroek: An integrated multimedia approach to cultural heritage, Proc. *ACM- Multimedia Information Retrieval 2002.*

[Smeulders 2000] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain: Content-based image retrieval at the end of the early years, *IEEE PAMI*, 1349–1380, 2000.

[Snoek 2005] C. Snoek, M. Worring, J. Geusebroek, D. Koelma, F. Seinstra, A. Smeulders: The semantic value chain. Accepted in *IEEE PAMI* based on MediaMill TRECVID 2004 Semantic Video Search.

[Tjong Kim Sang 2002] E.F. Tjong Kim Sang. Memory-based shallow parsing. *J. Machine Learning Research*, 2:559-594, 2002.

[Fergus 2003] R. Fergus, P. Perona, A. Zissermann: Object class recognition by unsupervised scale invariant learning. *CVPR 2003*, IEEE.

[Gevers 1999] T. Gevers, A Smeulders: Color based object recognition, *Patt Recognition*, 453-464, 1999.

[Geusebroek 2002] J.M. Geusebroek, R. van den Boomgaard, A.W.M. Smeulders, H. Gerards: Color invariants, *IEEE trPAMI*, 2002 1338-1350.

[Geusebroek 2005] J.M. Geusebroek, G.J Burghouts, A.W.M. Smeulders: The ALOI data set, *Int. J. Comp. Vis,* 2005 103-112.

[Geusebroek 2005] J.M. Geusebroek, A.W.M. Smeulders: A six stimulus theory for stochastic texture, *Int. J. Comp. Vis, 2005,* 7-16.

[Hauptmann 2004] A.G. Hauptmann, Towards a Large Scale Concept Ontology for Broadcast Video. 3rd CIVR, Dublin, 674-675, LNCS 3115 Springer 2004

[Hollink 2004] L. Hollink, G. Schreiber, B. Wielinga, M. Worring. Classification of User Image Descriptions. *Int. J. Human-Com Studies*, 2004.

[Hollink 2004] L. Hollink, G. Nguyen, D. Koelma, G, Schreiber, M. Worring. User Strategies in Video Retrieval. *CIVR 2004*, Dublin.

[Jackson and Moulinier 2002] P. Jackson I. Moulinier. *Natural Language Processing for Online Applications*. John Benjamins, 2002.

[Jijkoun 2004] V. Jijkoun, M. de Rijke, J. Mur. Information Extraction for Question Answering: *COLING 2004)*, 2004.

[Lowe 2004] D.G. Lowe, Distinctive image features from scale-invariant keypoints *Int. J. Comp. Vision,* 60, 2 (2004), 91-110.

[Maillot 2004] N. Maillot, M. Thonnat, A. Boucher. Towards ontology-based cognitive vision. *MVA Journal*, 16(1):33–40, 2004.

[Nguyen] Hieu Tat Nguyen, A.W.M. Smeulders: Everything gets better all the time apart from the amount of data. CIVR 2004, Dublin.

[NIST] TREC Video retrieval evaluation, 2001-2004.

[Mikolajczyk 2004] K. Mikolajczyk, C. Schmid: Scale and affine invariant interest point detectors. *Nt. Journ. Comp. Vis* 63 – 86, 2004 .

[Schober 2004] J-P. Schober, T. Hermes, O. Herzog: Content-based Image Retrieval by Ontology-based Object Recognition. In: V. Haarslev, C. Lutz, R. Möller (eds.), *Proc. Workshop Applications of Description Logics*, 2004.

[Sclaroff 1998] M. La Cascia, S. Sethi, S. Sclaroff: Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web. *Workshop on CB Access of Image and Video Libraries*, 1998

[Sigurbjörnsson 2004] B. Sigurbjörnsson, J. Kamps, and M. de Rijke, Processing Content-Oriented XPath Queries. *CIKM 2004*, 371-380

[Smeulders 2004] A.W.M. Smeulders, F. de Jong, M. Worring: Multimedia information technology for indexing *EU-conf video archives.* Proceedings still to appear.

[Stein 2003] G. Stein et al. Enabling video annotation using a semantic database extended with visual knowledge, *ICME 2003*.

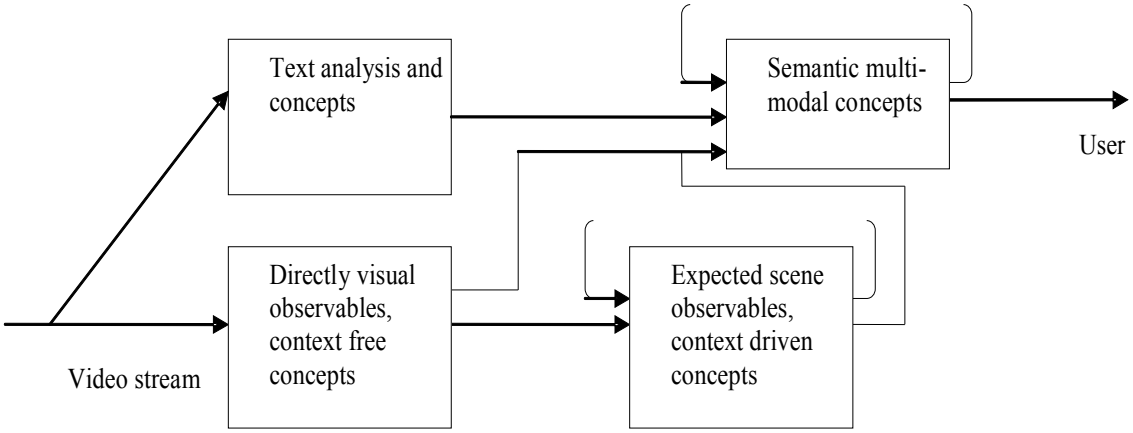

Figure 1: System layout of a video analysis system, integrating directly observables [Geusebroek 2005], with natural language analysis by unsupervised learning of paraphrases of text triggers of visual observables [Tjong Kim Sang 2002], machine learned semantic multi-modal concepts [Snoek 2005], and visual concepts as expected from the scene [VanGemert 2005].

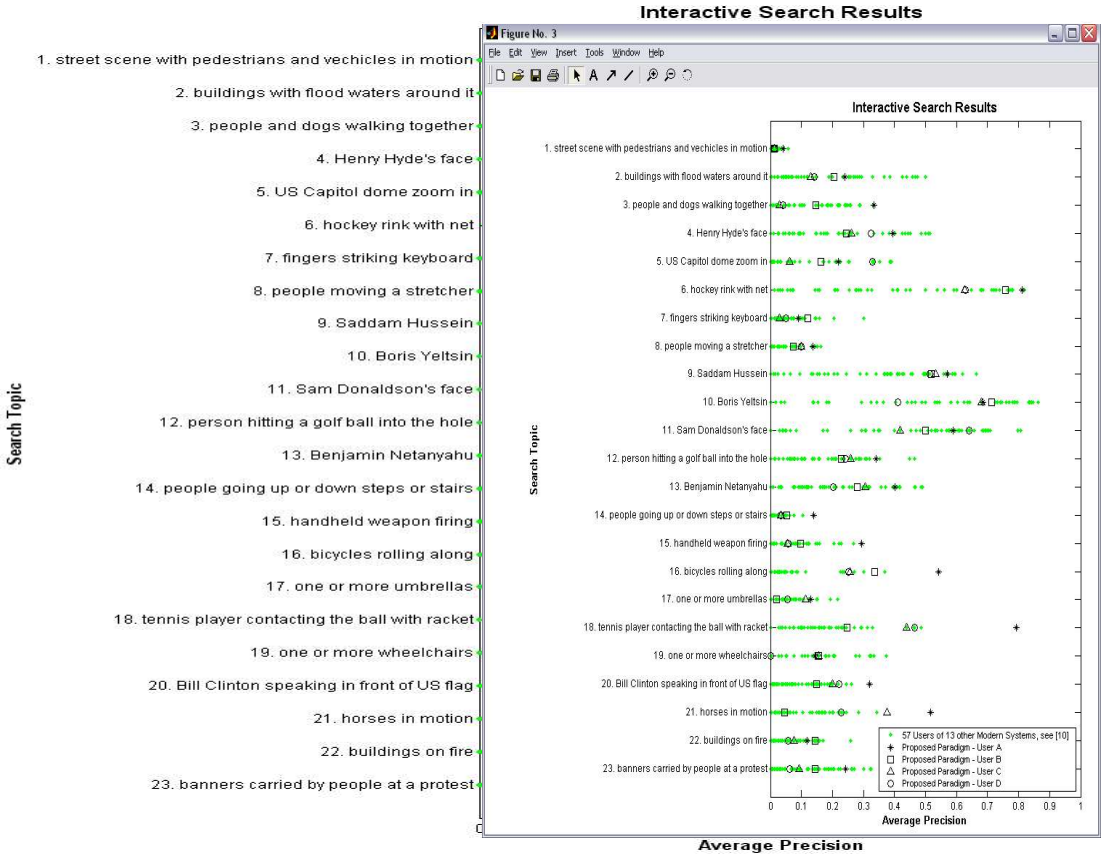

Table 1: Performance of the interactive video search engine in terms of average precision relative to the other participants in the 2004 edition of the TREC video competition [Snoek 2005]. Some questions like Boris Yeltsin rely on text processing almost entirely, whereas horses in motion – only indirectly present in our list of concepts – still performs well. Some concepts like pedestrians in motion does not get a reasonable response from any of the systems. Where the number of concepts in the thesaurus is limited still, these figures indicate that automated annotation and interactive search is already helpful in the search for background footage.