ON THE SURPLUS VALUE OF SEMANTIC VIDEO ANALYSIS BEYOND THE KEY FRAME

Cees G.M. Snoek, Marcel Worring, Jan-Mark Geusebroek, Dennis Koelma, and Frank J. Seinstra

ISLA, Informatics Institute, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
{cgmsnoek, worring, mark, koelma, fjseins}@science.uva.nl

Abstract

Typical semantic video analysis methods aim for classification of camera shots based on extracted features from a single key frame only. In this paper, we sketch a video analysis scenario and evaluate the benefit of analysis beyond the key frame for semantic concept detection performance. We developed detectors for a lexicon of 26 concepts, and evaluated their performance on 120 hours of video data. Results show that, on average, detection performance can increase with almost 40% when the analysis method takes more visual content into account.

1. INTRODUCTION

Methods for semantic video classification are evolving to a mature level. This is due to improved feature analysis methods [9], embedding of the problem into machine learning frameworks [4], and availability of large annotated data sets [5].

Despite these improvements, most methods still suffer from an often overlooked problem. Typical semantic video indexing approaches analyze a video at the granularity of a camera shot and try to predict its content based on extracted features, e.g. [3, 4, 12]. These approaches typically represent a shot by a single image only. These so called key frame based video analysis methods are thus deliberately ignoring a large amount of visual information.

Originally, the intention of using key frames was to aid in abstraction of visual content [1, 12]. In [1], a video is first segmented into shots, from every shot a representative frame is then chosen as key frame. By filtering out repetitive and uniform colored frames, the authors construct a visual table of contents. Motivated by a lack of processing power to entirely analyze large video archives, key frame-based analysis became a popular way for content-based video classification also [3, 4, 12].

While key frames are indeed valuable for quick browsing of the content of a video, they are not necessarily as suited for semantic classification. Moreover, while processing time was an obvious issue in the previous century, contemporary computing architectures allow for massive processing. Hence, for semantic video indexing there is no longer a need to restrict analysis to the level of key frames. In this paper we evaluate the use of key frames in a semantic video analysis scenario. Several of such approaches exist, e.g. [3, 4, 12]. We do not aim to judge specific semantic video analysis methods. Instead, we focus on the surplus value of analyzing video beyond the key frame.

The organization of the remainder of this paper is as follows. First, we outline a semantic video analysis scenario in Section 2. In Section 3, we describe the experimental setup in which the sketched method is evaluated. We discuss results in Section 4.

2. SEMANTIC VIDEO ANALYSIS SCENARIO

2.1. Weak Labeled Segmentation

Our scenario for semantic video analysis starts with weak segmentation of a single image frame, see also [10]. Ideally, a segmentation method should result in a semantically relevant partitioning of the image frame f, i.e. a strong segmentation. However, weak segmentation, where f is partitioned into internally homogenous regions based on some set of visual feature detectors, is often the best one can hope for [9]. We aim for weak segmentation.

Invariance was identified in [9] as a crucial aspect of a visual feature detector, e.g. to design features which limit the influence of accidental recording circumstances. As the conditions under which semantic concepts appear in large multimedia archives may vary greatly, we use invariant visual features to arrive at weak segmentation.

The procedure we adhere to computes per pixel a number of invariant visual features in vector $\vec{u}$. This vector then serves as the input for a multi-class supervised machine learner. This learner labels each pixel with one of the regional visual concepts defined in a visual concept lexicon Λ_V. For classification, the learner exploits an annotated training set. This pixel-wise classification of $\vec{u}$ results

This research is sponsored by the BSIK MultimediaN project.

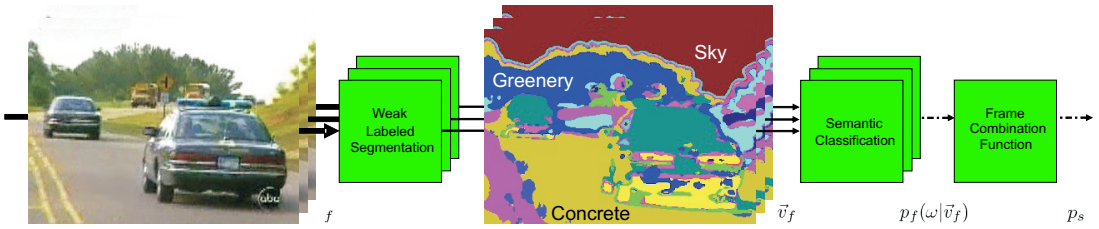

Fig. 1. Semantic video analysis scenario: from weak labeled segmentation to semantic classification of individual frames, and finally frame combination to generate a semantic index at shot level.

in a weak labeled segmentation of an image f in terms of regional visual concepts from Λ_V.

We use Gaussian color measurements [2] to obtain $\vec{u}$ for weak segmentation. We decorrelate RGB color values by linear transformation to the opponent color system [2]. Smoothing the values with a Gaussian filter suppresses acquisition and compression noise. The size of the Gaussian filters is varied to obtain a color representation that is compatible with variations in the target object size. Normalizing each opponent color value by its intensity suppresses global intensity variations. This results in two chromaticity values per color pixel. Furthermore, we obtain rotationally invariant features by taking Gaussian derivative filters and combining the responses into two chromatic gradients. The seven measurements in total, each calculated over three scales, yield a 21 dimensional feature vector $\vec{u}$ per pixel.

2.2. Semantic Classification

After weak labeled segmentation, the video analysis method uses the percentage of pixels associated to each of the regional visual concepts in Λ_V as a feature vector $\vec{v}_f$. Based on the distribution of regional visual concepts, a global classification of the f, in terms of semantic concepts ω from Λ_S, is made. To arrive at this concept classification, $\vec{v}_f$ forms the input for a supervised machine learning approach that associates a probability $p_f(\omega|\vec{v}_f)$ for all ω in Λ_S. After weak labeled segmentation and semantic classification we have indexed an image with a concept and its probability, see Fig. 1 for an overview of our analysis method.

2.3. Frame Combination Function

Within the sketched analysis method, we can apply several frame combination functions to generate a probability for ω from Λ_S at shot level s based on the probabilities of frames in s. In literature, most frame combination functions are surpassed by applying the analysis on a single key frame only. To show the merit of analysis beyond the key frame we introduce a *shot-based* combination function. For this function we make no assumptions about the data nor the analysis method. Consequently, all frames are treated equally. We average the probability $p_f(\omega|\vec{v}_f)$ over all analyzed frames to obtain a classification at shot level, defined as:

$$p_s = \sum_{f}^{f_s} p_f(\omega|\vec{v}_f) \tag{1}$$

We introduce two baseline key frame-based methods to show the merit of a shot-based semantic classification, i.e. a *pessimistic* and *optimistic* baseline. We view both baselines as optimal boundaries of the spectrum of key frame selection methods. The pessimistic baseline assumes the key frames coincide with the frame having the lowest semantic concept probability in a shot, or the frame with the weakest segmentation. Therefore, we consider the pessimistic baseline a worst-case scenario. It is defined as:

$$p_s^- = \arg\min_{f \in f_s} p_f(\omega|\vec{v}_f) \tag{2}$$

In contrast, the optimistic baseline assumes that the key frames coincide with the frame having the highest semantic concept probability in a shot, or the frame with the best weak segmentation. We consider this baseline a best-case scenario. It is defined as:

$$p_s^+ = \arg\max_{f \in f_s} p_f(\omega|\vec{v}_f) \tag{3}$$

Compared to the two key frame baselines, a shot-based frame combination takes all visual information within a shot into account. This should yield improved semantic concept detection performance.

3. EXPERIMENTAL SETUP

We performed an experiment as part of our participation in the 2004 NIST TRECVID video retrieval benchmark [5], to demonstrate that key frame-based analysis methods are inferior to methods that take more visual content into account.

3.1. Video Archive

The video archive of the 2004 TRECVID benchmark is composed of 184 hours of ABC World News Tonight and

CNN Headline News and recorded in MPEG-1 format. The development data contains approximately 120 hours. The test data contains the remaining 64 hours. Together with the video data, CLIPS-IMAG [7] provided a shot segmentation.

For our experiment, the 2004 TRECVID development data was split a priori into a non-overlapping training and validation set. The training set D contained 85% of the development data, the validation set V contained the remaining 15%. To each set, we alternatingly assign a proportional number of videos based on the broadcast date. For both sets this division assures maximum comparability.

3.2. Semantic Lexicons

In Section 2 we introduced two lexicons. The semantic concept lexicon, Λ_S, and the visual concept lexicon, Λ_V.

For Λ_S we define a lexicon of 26 semantic concepts. For all concepts considered, we annotated a ground truth [10]. The following concepts form the semantic concept lexicon:

- $\Lambda_S = \{$*airplane take off, American football, animal, baseball, basket scored, beach, bicycle, Bill Clinton, boat, car, cartoon, financial news anchor, golf, graphics, ice hockey, Madeleine Albright, overlayed text, people, people walking, physical violence, road, soccer, stock quotes, train, vegetation, weather news*$\}$;

Based on Λ_S we define the following set of regional visual concepts:

- $\Lambda_V = \{$*colored clothing, concrete, fire, graphic blue, graphic purple, graphic yellow, grassland, greenery, indoor sport court, red carpet, sand, skin, sky, smoke, snow/ice, tuxedo, water body, wood*$\}$;

This visual lexicon contains both general concepts, like grassland and water body, as well as specific concepts for this archive e.g. graphic blue and indoor sport court. As we use invariant visual features, only a few examples per visual concept class are needed, in practice less then 10 per class.

3.3. Supervised Machine Learner

A large variety of supervised machine learning approaches exists. For our purpose, the method of choice should handle typical problems related to semantic video analysis. Namely, it must learn from few examples, handle unbalanced data, and account for unknown or erroneous detected data. The Support Vector Machine (SVM) framework [11] is a solid choice in such a setting [4]. To obtain a probability from the SVM classifier, we convert its output using Platt's method [6]. To obtain optimal settings, $\vec{q}^*$ for an SVM classifier, parameter search on a large number of SVM parameter combinations must be applied [4]. The result of the parameter search is an optimized model, $p_f^*(\omega|\vec{v}_f, \vec{q}^*)$. The predefined training set D is used in combination with 3-fold cross validation to optimize SVM parameters for semantic concept detection.

3.4. Parallel Processing

Segmenting image frames into regional visual concepts at the granularity of a pixel is computationally intensive. We estimate that the processing of the entire TRECVID data set would have taken over 250 days on the fastest sequential machine available to us. As a first reduction of the analysis load, we analyze 1 out of 15 frames only. Where we note that the minimum duration of a shot is 60 frames. For the remaining image processing effort we apply the Parallel-Horus software architecture [8]. This architecture, consisting of a large collection of low-level image processing primitives, allows the programmer to write *fully sequential* applications for efficient parallel execution on homogeneous clusters of machines. This has a great impact on processing opportunities. Application of Parallel-Horus, in combination with a distributed Beowulf cluster consisting of 200 dual 1-Ghz Pentium-III CPUs, reduced the processing time to less than 60 hours [8].

4. RESULTS

We evaluated concept detection performance on lexicon Λ_S using the scenario sketched in Section 2. All shots in V are ranked according to the frame combination functions defined in (1), (2), and (3), using the parameter optimization discussed in Section 3.3 for all concepts in Λ_S.

4.1. Evaluation Criteria

The average precision, AP, is a single-valued measure that corresponds to the area under a recall-precision curve. This value is the average of the precision value obtained after each relevant shot is retrieved. Let $\mathcal{L}^j = \{l_1, l_2, \ldots, l_j\}$ be a ranked version of the answer set $\mathcal{A}$. At any given rank j let $\mathcal{R} \cap \mathcal{L}^j$ be the number of relevant shots in the top j of $\mathcal{L}$, where $\mathcal{R}$ the total number of relevant shots. Then AP of $\mathcal{L}^j$ is defined as:

$$AP(\mathcal{L}^j) = \frac{1}{\mathcal{R}} \sum_{j=1}^{\mathcal{A}} \frac{\mathcal{R} \cap \mathcal{L}^j}{j} \lambda(l_j) \qquad (4)$$

where $\lambda(l_j) = 1$ if $l_j \in \mathcal{R}$ and 0 otherwise. As the denominator j and the value of $\lambda(l_j)$ are crucial in determining AP, it can be understood that this metric favors highly ranked relevant shots. Based on the AP of two ranked answer sets, $\mathcal{L}_1^j$ and $\mathcal{L}_2^j$ we define the surplus value, SV, of $\mathcal{L}_2^j$ over $\mathcal{L}_1^j$ as:

$$SV(\mathcal{L}_1^j, \mathcal{L}_2^j) = \frac{AP(\mathcal{L}_2^j) - AP(\mathcal{L}_1^j)}{AP(\mathcal{L}_1^j)} * 100\% \qquad (5)$$

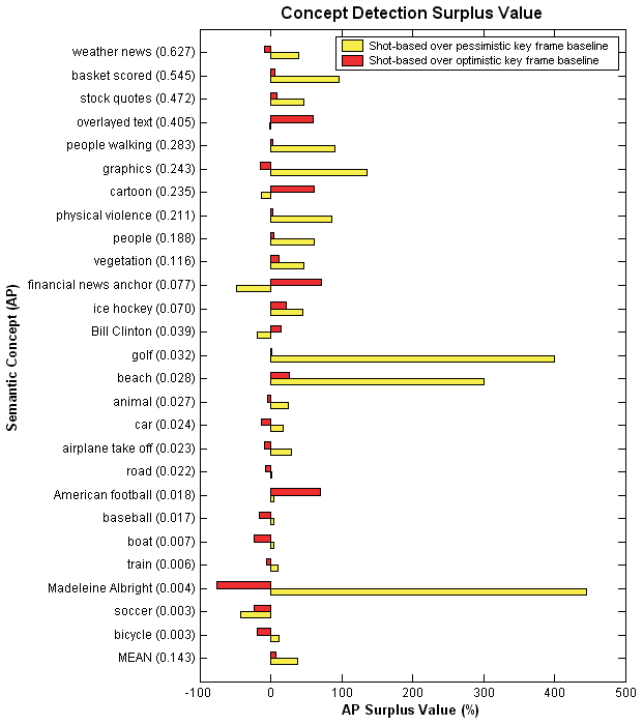

Fig. 2. Surplus value of semantic video analysis beyond the key frame.

We use the SV to compare shot-based combination with the two key frame baseline methods.

4.2. Concept Detection Surplus Value

We report the SV of shot-based analysis over the two key frame-based pessimistic and optimistic baseline methods on semantic lexicon Λ_S in Fig. 2. The mean of the SP over all 26 concepts shows that shot-based analysis outperforms both baselines. The surplus value of shot-based over the pessimistic baseline is 38.8%. Compared to the optimistic baseline, the surplus value is 7.5%. The SV over the different concepts varies. Compared to the pessimistic key frame baseline, a shot-based method is almost always better. For some concepts the SV reaches over 300%. Compared to the optimistic key frame baseline, the shot-based method performs better for most of the concepts. However, we observe that for concepts that are hard to detect based on visual analysis only, i.e. concepts that obtain a very low AP, an optimistic key frame selection method obtains better results. This suggests that for these concepts it is better to rely on one positive weak frame segmentation, then to average over multiple frames. Despite the low AP value for most concepts, the reported results are considered state-of-the-art performance within the TRECVID benchmark [10].

5. CONCLUSIONS AND FUTURE WORK

Semantic video classification methods should not focus their analysis on key frames only. In this paper, we demonstrated that a semantic video analysis method that considers more visual content obtains higher performance over key frame-based methods. The surplus value can range from 7.5% to 38.8%. Regions convey more semantic information than pixels. For future research, we therefore aim to extend the visual analysis method from pixel-based to region-based. Then, inclusion of spatial position will provide better performance. Furthermore, by considering all frames in the analysis, we can analyze temporal behavior of segmented regions. This will boost performance further.

6. ACKNOWLEDGEMENTS

The authors acknowledge Kees Verstoep from the Vrije Universiteit Amsterdam for DAS-2 support.

7. REFERENCES

[1] N. Dimitrova et al. Video keyframe extraction and filtering: a keyframe is not a keyframe to everyone. In *Int'l Conf. Inform. Knowledge Manage.*, pages 113–120, Las Vegas, USA, 1997.

[2] J. Geusebroek et al. Color invariance. *IEEE Trans. PAMI*, 23(12):1338–1350, 2001.

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

[4] M. Naphade. On supervision and statistical learning for semantic multimedia analysis. *J. Visual Commun. Image Representation*, 15(3):348–369, 2004.

[5] NIST. TREC Video Retrieval Evaluation, 2004. `http://www-nlpir.nist.gov/projects/trecvid/`.

[6] 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.

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

[8] F. Seinstra et al. User transparent parallel processing of the 2004 NIST TRECVID data set. In *Int'l Parallel Distrib. Processing Symposium*, Denver, USA, 2005.

[9] A. Smeulders et al. Content based image retrieval at the end of the early years. *IEEE Trans. PAMI*, 22(12):1349–1380, 2000.

[10] C. Snoek et al. The MediaMill TRECVID 2004 semantic video search engine. In *TRECVID Workshop*, Gaithersburg, USA, 2004.

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

[12] H.-J. Zhang et al. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30(4):643–658, 1997.