Proc. Fourth IEEE International Conference on Semantic Computing (ICSC 2010), Pittsburgh, PA, USA, September 2010, pp. 141-148.

On the use of visual soft semantics for video temporal decomposition to scenes

Vasileios Mezaris, Panagiotis Sidiropoulos, Anastasios Dimou, Ioannis Kompatsiaris
Informatics and Telematics Institute
Centre for Research and Technology Hellas
6th Km Charilaou-Thermi Road, Thermi 57001, Greece
Email: {bmezaris, psid, dimou, ikom}@iti.gr

***Abstract*—This work examines the possibility of exploiting, for the purpose of video segmentation to scenes, semantic information coming from the analysis of the visual modality. This information, in contrast to the low-level visual features typically used in previous approaches, is obtained by application of trained visual concept detectors such as those developed and evaluated as part of the TRECVID High-Level Feature Extraction Task. A large number of non-binary detectors is used for defining a high-dimensional semantic space. In this space, each shot is represented by the vector of detector confidence scores, and the similarity of two shots is evaluated by defining an appropriate shot semantic similarity measure. Evaluation of the proposed approach is performed on two test datasets, using baseline concept detectors trained on a dataset completely different from the test ones. The results show that the use of such semantic information, which we term "visual soft semantics", contributes to improved video decomposition to scenes.**

***Keywords*-video segmentation; scenes; visual soft semantics;**

I. Introduction

Video temporal decomposition into elementary semantic units is a prerequisite for a wide range of video processing and manipulation tasks, such as further semantic video analysis and classification [1], indexing, browsing, etc. One of the most important and commonly used elementary semantic units of video is the scene, which is often defined as a Logical Story Unit (LSU) [2], i.e. a series of temporally contiguous shots characterized by overlapping links that connect shots with similar content. Although different definitions of a "scene" exist, it is generally accepted that a scene needs to be a semantically and temporally coherent piece of video that is long enough to be meaningful on its own, i.e. to convey a story, in contrast to e.g. shots.

Previous approaches to scene segmentation focused on exploiting low-level visual or audio features for grouping similar shots into scenes, e.g. [2], [3]. In [3] in particular, the Scene Transition Graph (STG) was proposed; this method exploits the visual similarity between key-frames of video shots to construct a connected graph, whose cut-edges constitute the set of scene boundaries. More recent works present alternative schemes for evaluating shot similarities and combine low-level visual features with audio information. The latter includes low-level features (e.g. volume, sub-band energy, spectral and cepstral flux) as well as mid-level information (e.g. audio classification to silence, speech, music) and textual transcripts coming from automatic speech recognition [4], [5], [6].

Although low-level audiovisual information is useful for evaluating the similarity of shots for the purpose of grouping them, there is a gap between the similarities that can be revealed by examining low-level properties of the audiovisual signal and the semantic coherence that is desired of a scene. To bridge this gap, the use of a number of high-level audio events in combination with other semantic audio analysis results (e.g. speaker segmentation) and low-level visual information, was proposed in [7]. An appropriate method based on the extension of the STG approach was developed for jointly considering the above information, and its evaluation showed that the use of semantic audio information contributes to significantly improved results, compared to using low-level audiovisual information alone.

In this work, we examine the possibility of exploiting, for the purpose of video segmentation to scenes, semantic information coming from the analysis of the visual modality. This information, in contrast to the low-level visual features typically used in previous approaches, is obtained by application of trained visual concept detectors, such as those developed and evaluated as part of the TRECVID High-Level Feature Extraction Task [8]. A large number of non-binary detectors is used for defining a high-dimensional semantic space. In this space, each shot is represented by the vector of detector confidence scores (similar to the "model vector" of [9]), each score being a real number in the range 0 to 1. The similarity of two shots is evaluated by defining an appropriate shot semantic similarity measure. Similar use of visual concept detectors is made in certain image/video retrieval tasks, e.g. [10], although generally in concept-based (or semantic) retrieval the emphasis is on selecting a small subset of concepts that are relevant to a given query, rather than on exploring the entire semantic space [11], [12]. We show that the use of such semantic information, which we term "visual soft semantics" to indicate that it encompasses uncertainty about the semantics of each piece of visual information, contributes to improved video decomposition. This improvement is demonstrated using baseline concept detectors (i.e. detectors inferior to the current state-of-the-

art) trained on a dataset completely different from the test ones, thus highlighting the usefulness of large numbers of realistic visual concept detectors in this task.

The rest of the paper is organized as follows: an overview of the proposed approach is presented in section II. Visual soft semantics extraction and use for shot representation and similarity evaluation are discussed in section III, while a video temporal segmentation algorithm that exploits visual soft semantics is presented in section IV. Experimental results in two different datasets are presented in section V and conclusions are drawn in Section VI.

II. Overview

Temporal video segmentation to scenes is performed under the proposed approach by clustering video shots to temporally contiguous clusters, as is typically the case in the relevant literature. Thus, the process starts with application of the shot segmentation algorithms of [13], [14] (for abrupt and gradual transition detection, respectively), which generate a decomposition S of the video to visual shots,

$$S = \{s_i\}_{i=1}^{I} \quad (1)$$

Subsequently, as illustrated in Fig. 1, previously trained visual concept detectors are used for extracting soft semantic information from the visual content. This information is used, possibly together with low-level visual features, for representing the shots. The resulting shot representations serve as input to a temporal segmentation algorithm, which performs the detection of the scene boundaries.

III. Visual Soft Semantics

A. Hard versus soft semantics

Semantics can be captured and represented in a variety of ways, depending on several factors such as the content in question (e.g. text, images/videos, medical data), the target use of them (e.g. information integration, content retrieval), and the specific techniques employed for their extraction from the content (e.g. crisp clustering, statistical learning). In [15], with the Semantic Web in mind, three different forms of semantics are identified, namely "implicit", "formal", and "powerful" (or "soft"), and the importance of powerful semantics is advocated. In the context of audiovisual content analysis, on the other hand, one can easily identify two broad classes of semantic information: that representing binary relations between content and concepts, which we term "hard semantic information" or "hard semantics" for short (e.g. "image x depicts B. Clinton"), and that encompassing uncertainty about the content-concept relation, which we term "soft semantics" (e.g. "image x depicts B. Clinton with 0.7 confidence"). The latter closely relates to the powerful semantics of [15], in that the notion of uncertainty is central to both definitions. Following the discussion in the aforementioned work on the importance of powerful semantics, as well as considering the particular limitations in state-of-the-art semantic information extraction from visual content, we concentrate in this work on examining the use of uncertain semantic information coming from the visual modality ("visual soft semantics") in video temporal decomposition.

B. Semantic information extraction from visual content

The automatic association of visual content with semantic concepts in this work is based on a relatively simple (baseline) approach, which revolves around treating each concept separately from all others, and using late fusion for combining concept detection results attained for a single concept with the use of different sets of visual descriptors.

Starting with the semantic concepts that are used, these are the 101 concepts defined for the TRECVID 2005 dataset as part of the Mediamill challenge [16]. For training detectors for these concepts, the TRECVID 2005 training dataset and the corresponding Mediamill ground truth annotations are employed.

Using the aforementioned concepts and annotated dataset, a concept detector is trained for each concept separately. For this, a set of MPEG-7 features (color structure, color layout, edge histogram, homogeneous texture and scalable color) [17] are initially extracted from the keyframes of the video dataset (one keyframe per shot) and are concatenated to form a single MPEG-7 feature vector of the keyframe. In parallel to this, a Bag-of-Words (BoW) feature vector is also calculated for each keyframe, following the extraction of SIFT descriptors and the construction of a small vocabulary of 100 visual words [18]. Subsequently, a two stage classification process is realized by training for each concept two Support Vector Machine (SVM) classifiers (one using the MPEG-7 feature vector and the other using the BoW one) and using their output for training a third SVM classifier that realizes late fusion. The output of each of the aforementioned SVMs is a number in the continuous range $[0, 1]$, expressing the Degree of Confidence (DoC) that the keyframe relates to the corresponding concept. The two-stage trained concept detectors are evaluated in the "testing" part of the TRECVID 2005 dataset and suitable performance measures (Average Precision (AP), Delta-Average Precision (ΔAP) [19]) are calculated for each. The employed concept detection technique was shown in our TRECVID 2008 experiments [20] to rank close to the median, thus it generates moderately accurate concept detectors compared to the current state-of-the-art.

The trained concept detectors resulting from the above described process can subsequently be used on any dataset, following feature extraction, for estimating a DoC value in the [0, 1] range for every given keyframe-concept pair.

C. Shot representation and similarity evaluation

Application of J different trained visual concept detectors on a keyframe f results in J DoC values, which can be

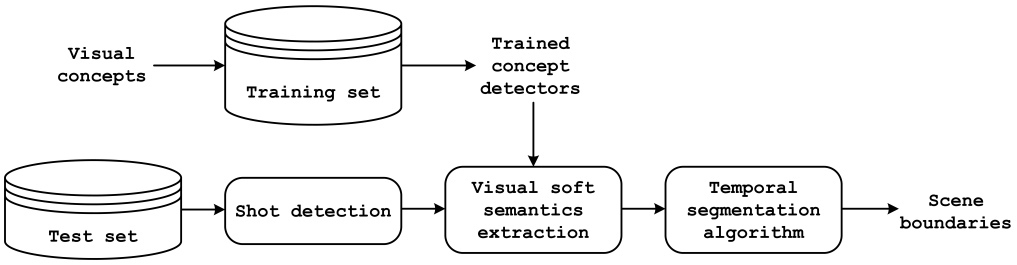


Figure 1. Overview of the proposed approach for video temporal decomposition to scenes.

expressed as a vector $c(f)$,

$$c(f) = [c_1(f), c_2(f), ..., c_J(f)] \quad (2)$$

This vector essentially represents keyframe f in the semantic space defined by the J concepts. Subsequently, in order to take into account the results of concept detection in more than one keyframes per shot, the shot representation vector $c(s)$ is defined as:

$$c(s) = [c_1(s), c_2(s), ..., c_J(s)] \quad (3)$$

$$c_j(s) = \max_{f \in s}\{c_j(f)\} \quad (4)$$

The rationale behind this choice is that, for the purpose of scene segmentation, it is most useful to know which concepts are more likely to be visible in at least part of the shot.

The calculation of $c(s)$ is followed by a normalization step. Similarly to audio events, discussed in [7], different visual concepts may have different frequency of appearance in a given video (i.e. some concepts are more rare than others). Because of this and also of the specifics of each trained detector, the detectors may consistently produce lower- or higher-than-average DoC values. This lack of homogeneity can affect the evaluation and comparison of differences in the semantic space that the detectors define, e.g. by minimizing the impact of detectors that consistently produce low DoC values. To alleviate this, the normalization of values $c_j(s)$ is proposed, and in this work a very simple normalization approach is adopted. Specifically, the elements of the normalized shot representation vector $\tilde{c}(s)$ are estimated as:

$$\tilde{c}_j(s) = \frac{c_j(s)}{maxc_j^S} \quad (5)$$

where $maxc_j^S$ is the maximum value of the j-th concept detector in all shots of the examined video.

Following normalization, the definition of a shot similarity measure is based on the requirement that not only the difference of values $\tilde{c}_j(s)$ between two shots, but also the absolute values $\tilde{c}_j(s_i)$ and $\tilde{c}_j(s_k)$ themselves, should affect shot similarity. The rationale behind this is that, for the j-th detector, two shots receiving similarly high confidence values is a strong indication of their semantic similarity (i.e. they are both likely to depict the j-th concept). On the contrary, the same shots receiving similarly low confidence values is an indication neither in favor nor against their semantic similarity; it merely suggests that the j-th concept (out of a large number J of concepts) is not depicted in either of the two shots. The commonly used Minkowski distance does not satisfy the above requirement, since it depends only on the difference of the confidence values. Instead of it, a variation of the Chi-test distance, that was shown to be useful when considering audio events [7], is employed in this work. Thus, the distance D of $\tilde{c}(s_i)$ and $\tilde{c}(s_k)$ is defined as:

$$D(\tilde{c}(s_i), \tilde{c}(s_k)) = \sqrt{\sum_{j=1}^{J} \frac{(\tilde{c}_j(s_i) - \tilde{c}_j(s_k))^2}{\tilde{c}_j(s_i) + \tilde{c}_j(s_k)}} \quad (6)$$

This similarity measure is used in the temporal segmentation algorithm.

IV. Temporal Segmentation Algorithm

Temporal segmentation to scenes is performed using an extension of the Scene Transition Graph (STG) algorithm of [3]. The main characteristics of the extended algorithm include i) the definition of a STG that exploits visual soft semantics rather than low-level visual features, to allow for shot similarity evaluation at a more semantic level, and ii) the construction of multiple STGs and their combination in a probabilistic merging process, to reduce the dependence of the original STG on the values of its construction parameters.

Depending on the information used for evaluating shot similarity, i.e. low-level visual features or visual soft semantics, the proposed approach distinguishes between two types of STGs: the VSTG and the Visual Concept STG, respectively. For constructing an STG of the first type, the original method of [3] is adopted: HSV histograms of the keyframes are extracted and used for clustering visually similar shots; then, a graph is formed with its nodes representing the shot clusters and directed edges between the nodes expressing the temporal succession of the shots that are included in the clusters; finally, "cut-edges" (:edges that, if removed, result in two disconnected graphs) are identified and are declared scene boundaries. The Visual

Concept STG, on the other hand, is defined in this work in a fashion similar to the VSTG, using however $\tilde{c}(s)$ and $D(\tilde{c}(s_i), \tilde{c}(s_k))$ (Eqs. (2)-(6)) for representing the shots and for evaluating their similarity, instead of HSV histograms and the Euclidean distance used in VSTG.

The construction of an STG of any one of the aforementioned types requires the selection and use of a number of parameters (similarity threshold, temporal distance threshold), which are typically selected heuristically. In order to alleviate the need for this, we introduced in [21] a probabilistic STG merging approach that combines multiple STGs and simultaneously reduces the dependency of the combination on individual STG construction parameters. Following this approach, in this work multiple ($\mathcal{P}$; $\mathcal{P} \gg 1$) VSTGs are created, each using a different randomly selected set of parameter values. Then, the fraction p_i^v of VSTGs that identify the boundary between shots s_i and s_{i+1} as a scene boundary (i.e. the number of such VSTGs, divided by the total number of generated VSTGs) is calculated and used as a measure of our confidence on this being a scene boundary, based on low-level visual information. The same procedure is followed using $\mathcal{P}$ Visual Concept STGs, resulting in confidence values p_i^s. Subsequently, these confidence values are linearly combined to result in a final confidence value p_i:

$$p_i = V \cdot p_i^v + (1 - V) \cdot p_i^s \tag{7}$$

In the above formula, V is a global parameter that controls the relative weight of the VSTGs and Visual Concept STGs in the scene boundary estimation. Finally, all shot boundaries (s_i, s_{i+1}) for which p_i exceeds a threshold,

$$\Gamma = \{(s_i, s_{i+1}) | p_i > T\} \tag{8}$$

form the set Γ of scene boundaries estimated by the proposed approach. In the remainder of the paper, the terms "VSTG" and "Visual Concept STG" are used for denoting the overall approach of Eqs. (7)-(8) when V in Eq. (7) is equal to 1 and 0, respectively, rather than a single one of the $\mathcal{P}$ STGs of each type that are constructed as part of the proposed approach.

V. Experimental Results

A. Datasets and experimental setup

For experimentation, two test-sets were used. The first one is made of 7 documentary films (229 minutes in total) from the collection of the Netherlands Institute for Sound & Vision[1]. The second one is made of three movies (330 minutes in total). Application of the shot segmentation algorithms of [13], [14] (for abrupt and gradual transition detection, respectively) to these test-sets resulted in 1444 and 3638 shots; manual grouping of them to scenes resulted in 237 and 177 ground truth scenes. For each of the two datasets, one additional video of the same gender (one documentary, one movie) was processed in the same way (shot segmentation, manual grouping of the shots to scenes) and was used for automatically adjusting the parameters of the algorithm (T, V) in some of the reported experiments.

For evaluating the results of the scene segmentation experiments, the Coverage (C), Overflow (O) and F-score (F) measures were employed. Coverage and Overflow were proposed in [22] for scene segmentation evaluation; Coverage measures to what extent frames belonging to the same scene are correctly grouped together, while Overflow evaluates the quantity of frames that, although not belonging to the same scene, are erroneously grouped together (see [22] for complete definitions). The optimal values for Coverage and Overflow are 100% and 0% respectively. The F-score is defined in this work as the harmonic mean of C and $1 - O$, to combine Coverage and Overflow in a single measure,

$$F = \frac{2C(1 - O)}{C + (1 - O)} \tag{9}$$

were $1 - O$ is used above instead of O to account for 0 being the optimal value of the latter, instead of 1.

A first set of experiments ($E1$) was carried out by setting $V = 1$ in Eq. (7), thus not using the concept detection results at all (only VSTG results are used). The resulting method essentially resembles the original STG method of [3], integrating however the technique introduced in section IV for reducing the influence of STG construction parameters to the final temporal video decomposition. For this experiment, three keyframes per shot were used. The number $\mathcal{P}$ of STGs constructed using randomly selected parameters was set to 1000; the reader is referred to [3] for a discussion on STG parameters. The value of threshold T of Eq. (8) was chosen by exhaustive search (with step 0.01 in the range $(0, 1)$) as the one that maximizes the F-score attained for the test dataset. The latter was done for calculating an upper bound for the performance of VSTG; experiments with automatic selection of the value of T are also reported in the sequel.

A second set of experiments ($E2$) was carried out by setting $V = 0$ in Eq. (7), thus studying the possibility of using only the visual soft semantics of section III for shot representation. During this set of experiments, the number J of concept detectors that were taken into account (Eq. (6)) was varied from 10 to 90 with a step of 10; using all 101 concept detectors was also examined. Assuming that, when selecting a subset of the available detectors, it makes sense to select the best J detectors out of all the available ones, two different "goodness" criteria were used for the detectors: Average Precision (AP) and Delta Average Precision (ΔAP) [19]. Both AP and ΔAP for the trained concept detectors were those calculated on the test portion of the TRECVID 2005 dataset. The value of T was chosen as in the first set of experiments, for the reasons discussed above. In this and all subsequent experiment sets, the same

[1] http://instituut.beeldengeluid.nl/

keyframes used in the first experiment set were used, and $\mathcal{P} = 1000$.

In a third set of experiments ($E3$), the experiments of the second set were repeated with $V \neq 0$; in this case, the values of both V and T were chosen by exhaustive search as the ones that maximize the F-score attained for the test dataset. This set of experiments aims at revealing the potential increase in performance when combining the VSTG and Visual Concept STG approaches, thus combining the low-level features typically used for shot representation with the representation based on visual soft semantics. Again, the performance scores reported for this experiment are only upper bounds.

A fourth set of experiments ($E4$) was carried out by repeating the third one, using however an automatically selected value of V; the latter was selected using an out-of-testset ground-truth-segmented video of the same gender as the test-set considered each time (see the beginning of section V) and Least Squares Estimation (LSE). The results of this set of experiments are directly comparable to those of the first and second sets, since in all three cases the only parameter value chosen a posteriori is T.

To examine the impact of also selecting the latter automatically, using the additional (out-of-testset) ground-truth-segmented video mentioned before, two more sets of experiments were run. These ($E5$ and $E6$) were repetitions of the first and fourth experiment set, respectively, with the value of T in both cases being automatically set to the value that is optimal (in terms of F-score) for segmenting the out-of-testset video. These fifth and sixth sets of experiments indicate the impact of automatically selecting T with the use of a small ground-truth-segmented corpus; the results of these two sets of experiments are directly comparable with each other, since all parameter values are in both cases selected automatically.

B. Results and discussion

The results of all the aforementioned experiments are presented in Table I for the Movie dataset and in Table II for the Documentary dataset. Since in all experiments parameters T and V were chosen (either by exhaustive search on the test set, or automatically using an out-of-testset video) so that they maximize an F-score, the discussion of the results will concentrate on the reported F-scores.

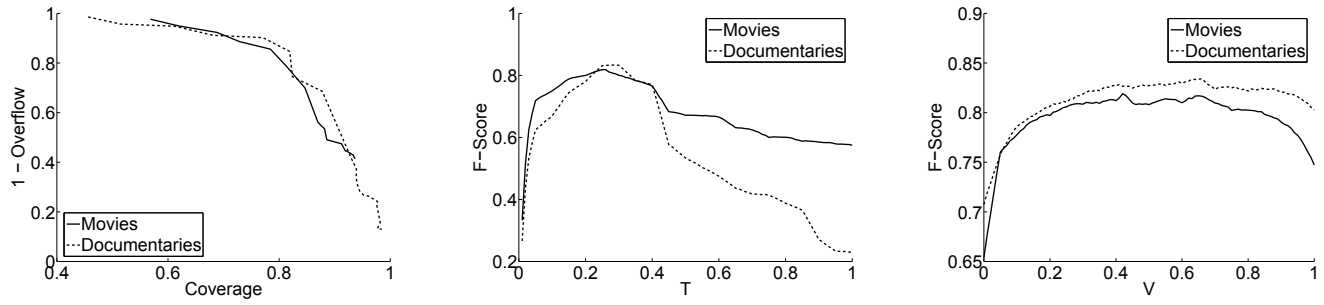
Starting with $E1$ and $E2$, it can be seen that using visual soft semantics alone does not produce improved results compared to the baseline VSTG; however, the results are in one case identical. The results of $E2$ vary significantly with the number of concept detectors J; regardless of whether AP of ΔAP is used as the selection criterion, in general the use of a larger number of detectors leads to improved results. A possible explanation of this is that, in the absence of low-level features, even "bad" concept detectors can contribute by producing similar responses for "similar" keyframes (if not semantically similar, at least visually similar), thus meaningfully increasing the dimensionality of the feature space in which the similarity of shots is examined.

When considering the combination of VSTG and Visual Concept STG ($E3$), though, the impact of "bad" concept detectors on the results is reversed: the best results are attained for J lower that 101. This can be attributed to the use of the low-level features in VSTG, which render "bad" detectors unnecessary. Thus, such detectors in this case seem to only introduce additional noise to the representation of the shots; this noise is responsible for the slight decline of the F-score when increasing the value of J beyond an optimal one. More specifically, in the Movie dataset the combination of VSTG and Visual Concept STG is shown to result in a maximum F-score of 81.91%, compared to 74.94% for the VSTG alone; the corresponding figures for the Documentary dataset are 83.42% and 80.66% respectively. It is important to observe, however, that the results of $E3$ are consistently better than those of both $E1$ and $E2$, regardless of the number of considered concepts. The clear superiority of the combination of VSTG and Visual Concept STG over the VSTG alone is maintained when the relevant significance of VSTG and Visual Concept STG in their combination (weight V) is determined automatically using an out-of-testset video ($E4$). In Fig. 2, the impact of parameters V and T is further illustrated, for the best-performing configuration of $E3$ ($J = 50$; concepts selected according to ΔAP), by varying one of these two parameters at each time. From this figure it is clear that T significantly affects performance, as expected, while for V there is a relatively broad range of values that result in only minor performance fluctuations.

Additional comments on the $E3$ and $E4$ results have to do with the concept detector "goodness" criteria and with the optimal value of V. Concerning the former, it can be seen that the use of ΔAP is advantageous over AP, since it i) leads to a higher maximum F-score, and ii) allows reaching this maximum with the use of a lower number of concepts. Concerning the optimal value of V, differences are observed among the two employed datasets: while in the Movie dataset the best results are obtained in most cases for $V \leq 0.5$, in the Documentary dataset $V > 0.6$ is typically required. This can be attributed to qualitative differences between the videos of each of the two datasets.

Reviewing the results of $E5$ and $E6$, where (besides V, where applicable) threshold T is also selected automatically using the out-of-testset video, is can be seen that there is a relatively small decline in all F-scores. However, the results of $E6$ are consistently better than those of $E5$, regardless of the number of considered concepts. Interestingly, the results of $E6$ are also consistently better than the results of $E1$, where the optimal value of T was used (T was determined in $E1$ by exhaustive search on the employed testset).

Finally, it should be emphasized that in our experiments the detectors are used in two completely different datasets



(a) (b) (c)

Figure 2. Impact of parameters T and V, for the best-performing configuration of experiment set $E3$ ($J = 50$; concepts selected according to ΔAP): (a) 1-Overflow versus Coverage when T varies from 0.05 to 1 (V=const), (b) F-score versus T (V=const), (c) F-score versus V (T=const).

than the TRECVID 2005 one, in which they were trained and evaluated according to AP or ΔAP. Thus, the contribution of visual soft semantics to improved video temporal decomposition that is reported in Tables I and II is attained in spite of the visual concept detectors being in general unreliable, a fact clearly documented in [19].

VI. Conclusion

In this work, the possibility of exploiting semantic information coming from the analysis of the visual modality for the purpose of video segmentation to scenes was examined. A high-dimensional semantic space was defined, with the use of trained visual concept detectors, and the shot representations in this space were used for grouping the shots to scenes. Experiments on two datasets, using realistic concept detectors, revealed the performance improvement that can be achieved by the introduction of visual soft semantics to a low-level-feature-based scene segmentation method.

Acknowledgment

This work was supported by the European Commission under contract FP7-248984 GLOCAL.

References

[1] G. Papadopoulos, A. Briassouli, V. Mezaris, I. Kompatsiaris, and M. Strintzis, "Statistical motion information extraction and representation for semantic video analysis," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 19, no. 10, pp. 1513–1528, October 2009.

[2] A. Hanjalic and R. L. Lagendijk, "Automated high-level movie segmentation for advanced video-retrieval systems," *IEEE Trans. On Circuits and Systems for Video Technology*, vol. 9, pp. 580–588, June 1999.

[3] M. Yeung, B.-L. Yeo, and B. Liu, "Segmentation of video by clustering and graph analysis," *Computer Vision and Image Understanding*, vol. 71, pp. 94–109, July 1998.

[4] A. Chianese, V. Moscato, A. Penta, and A. Picariello, "Scene detection using visual and audio attention," in *ACM Int. Conf. on Ambient Media and Systems*, Quebec, Canada, February 2008.

[5] K. Wilson and A. Divakaran, "Discriminative genre-independent audio-visual scene change detection," in *SPIE Conf. on Multimedia Content Access: Algorithms and Systems III*, vol. 7255, 2009.

[6] W. Jinqiao, D. Lingyu, L. Qingshan, L. Hanqing, and J. Jin, "A multimodal scheme for program segmentation and representation in broadcast video streams," *IEEE Trans. on Multimedia*, vol. 10, pp. 393–408, April 2008.

[7] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso, "On the use of audio events for improving video scene segmentation," in *Proc. WIAMIS*, April 2010.

[8] A. F. Smeaton, P. Over, and W. Kraaij, "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements," in *Multimedia Content Analysis, Theory and Applications*, A. Divakaran, Ed. Berlin: Springer Verlag, 2009, pp. 151–174.

[9] J. Smith, M. Naphade, and A. Natsev, "Multimedia semantic indexing using model vectors," in *Proc. IEEE ICME*, Baltimore, MD, USA, July 2003, pp. 445–448.

[10] N. Rasiwasia, P. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Trans. on Multimedia*, vol. 9, no. 5, pp. 923–938, 2007.

[11] J. Cao, H. Jing, C.-W. Ngo, and Y. Zhang, "Distribution-based concept selection for concept-based video retrieval," in *Proc. ACM Multimedia*, Beijing, China, October 2009, pp. 645–648.

[12] C. Snoek and M. Worring, "Concept-based video retrieval," *Foundations and Trends in Information Retrieval*, vol. 4, no. 2, p. 215322, 2009.

[13] G. Chavez, M. Cord, S. Philip-Foliguet, F. Precioso, and A. Araujo, "Robust scene cut detection by supervised learning," in *Proc. EUSIPCO*, Florence, Italy, September 2006.

[14] E. Tsamoura, V. Mezaris, and I. Kompatsiaris, "Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework," in *Proc. IEEE ICIP, MIR Workshop (ICIP-MIR 2008)*, San Diego, CA, USA, October 2008, pp. 45–48.

[15] A. Sheth, C. Ramakrishnan, and C. Thomas, "Semantics for the semantic web: the implicit, the formal and the powerful," *Int. Journal on Semantic Web and Information Systems*, vol. 1, pp. 1–18, 2005.

[16] C. Snoek, M. Worring, J. van Gemert, J.-M. Geusebroek, and A. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proc. ACM Multimedia*, Santa Barbara, USA, October 2006, pp. 421–430.

[17] B. Manjunath, J.-R. Ohm, V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 703–715, June 2001.

[18] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka, "Visual categorization with bags of keypoints," in *Proc. ECCV Int. Workshop on Statistical Learning in Computer Vision*, Prague, CZ, May 2004.

[19] J. Yang and A. Hauptmann, "(Un)Reliability of video concept detection," in *Proc. ACM CIVR*, Niagara Falls, Canada, July 2008.

[20] J. Molina, V. Mezaris, P. Villegas, and et. al., "MESH participation to TRECVID2008 HLFE," in *Proc. TRECVID 2008 Workshop*, Gaithersburg, MD, USA, November 2008.

[21] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, and I. Trancoso, "Multi-modal scene segmentation using scene transition graphs," in *Proc. ACM Multimedia*, Beijing, China, October 2009, pp. 665–668.

[22] J. Vendrig and M. Worring, "Systematic evaluation of logical story unit segmentation," *IEEE Trans. on Multimedia*, vol. 4, pp. 492–499, December 2002.

Table I
RESULTS OF VIDEO TEMPORAL DECOMPOSITION TO SCENES FOR THE MOVIE DATASET

$E1$: VSTG; optimal threshold T (selected by exhaustive search on the test dataset)

Concept Num. J	Coverage(%)	Overflow(%)	F-score(%)	VSTG weight V
-	71.44	21.19	74.94	1

$E2$: Visual Concept STG; optimal threshold T (selected by exhaustive search on the test dataset)

	Visual concepts selected according to their AP				Visual concepts selected according to their ΔAP			
Concept Num. J	Coverage(%)	Overflow(%)	F-score(%)	Weight V	Coverage(%)	Overflow(%)	F-score(%)	Weight V
10	79.74	47.44	63.36	0	78.8	45.31	64.57	0
20	78.34	43.03	65.97	0	77.71	38.22	68.84	0
30	76.84	40.57	67.02	0	80.16	45.19	65.10	0
40	74.37	40.68	66.00	0	76.36	38.76	67.97	0
50	71.32	37.26	66.76	0	73.95	40.23	66.11	0
60	70.86	32.84	68.96	0	72.67	36.73	67.65	0
70	67.09	27	69.92	0	70.89	27.68	71.60	0
80	73.56	31.8	70.78	0	70.14	28.06	71.03	0
90	**78.83**	**28.57**	**74.95**	0	75.08	28.96	73.00	0
101	74.73	27.3	73.70	0	**74.73**	**27.3**	**73.70**	0

$E3$: Combination of Visual Concept STG and VSTG; optimal threshold T and VSTG weight V (both selected by exhaustive search on the test dataset)

	Visual concepts selected according to their AP				Visual concepts selected according to their ΔAP			
Concept Num. J	Coverage(%)	Overflow(%)	F-score(%)	Weight V	Coverage(%)	Overflow(%)	F-score(%)	Weight V
10	79.06	18.15	80.43	0.52	77.8	17.31	80.17	0.53
20	78.39	17.43	80.43	0.52	76.42	16.03	80.02	0.64
30	77.51	16.24	80.51	0.44	77.98	15.75	80.99	0.65
40	80.17	18.38	80.89	0.54	79.17	15.9	81.56	0.51
50	79.67	17.83	80.90	0.68	**78.46**	**14.32**	**81.91**	0.42
60	79.43	18.00	80.69	0.66	78.43	14.45	81.84	0.39
70	**80.67**	**18.77**	**80.95**	0.48	77.67	14.6	81.35	0.63
80	77.94	16.95	80.41	0.61	75.28	13.91	80.32	0.46
90	76.30	15.16	80.34	0.52	76.07	14.76	80.39	0.41
101	74.82	15.22	79.49	0.57	74.82	15.22	79.49	0.57

$E4$: Combination of Visual Concept STG and VSTG; VSTG weight V selected automatically using an out-of-testset video; optimal threshold T (selected by exhaustive search on the test dataset)

	Visual concepts selected according to their AP				Visual concepts selected according to their ΔAP			
Concept Num. J	Coverage(%)	Overflow(%)	F-score(%)	Weight V	Coverage(%)	Overflow(%)	F-score(%)	Weight V
10	78.78	19.37	79.69	0.44	81.31	21.72	79.77	0.69
20	80.64	20.00	80.32	0.69	76.79	16.53	79.99	0.65
30	75.39	14.84	79.98	0.37	79.51	17.85	80.81	0.53
40	79.75	18.59	80.57	0.6	77.69	15.32	81.04	0.43
50	79.82	18.59	80.61	0.57	**78.25**	**14.46**	**81.73**	0.5
60	**79.43**	**18.00**	**80.69**	0.66	81.76	19.72	81.01	0.44
70	86.72	24.87	80.51	0.55	78.37	15.52	81.31	0.64
80	77.68	19.74	78.95	0.43	74.30	15.27	79.17	0.62
90	77.33	16.86	80.13	0.55	76.74	16.27	80.08	0.38
101	74.30	16.12	78.80	0.62	74.30	16.12	78.80	0.62

$E5$: VSTG; threshold T selected automatically using an out-of-testset video

Concept Num. J	Coverage(%)	Overflow(%)	F-score(%)	VSTG weight V
-	68.17	20.38	73.45	1

$E6$: Combination of Visual Concept STG and VSTG; threshold T and VSTG weight V selected automatically using an out-of-testset video

	Visual concepts selected according to their AP				Visual concepts selected according to their ΔAP			
Concept Num. J	Coverage(%)	Overflow(%)	F-score(%)	Weight V	Coverage(%)	Overflow(%)	F-score(%)	Weight V
10	76.28	16.89	79.55	0.44	81.31	21.72	79.77	0.69
20	80.18	19.72	80.23	0.69	75.57	16.08	79.53	0.65
30	72.15	10.93	79.72	0.37	80	18.64	80.67	0.53
40	81.37	21.04	80.15	0.6	74.75	12.59	80.59	0.43
50	**80.76**	**19.61**	**80.57**	0.57	**79.01**	**16.24**	**81.32**	0.5
60	77.66	16.49	80.48	0.66	82.18	20.66	80.74	0.44
70	86.72	24.87	80.51	0.55	78.12	15.3	81.28	0.64
80	75.06	17.24	78.72	0.43	69.27	9.53	78.46	0.62
90	81.17	22.35	79.37	0.55	78.5	19.07	79.70	0.38
101	74.68	16.62	78.79	0.62	74.68	16.62	78.79	0.62

Table II
RESULTS OF VIDEO TEMPORAL DECOMPOSITION TO SCENES FOR THE DOCUMENTARY DATASET

$E1$: VSTG; optimal threshold T (selected by exhaustive search on the test dataset)

Concept Num. J	Coverage(%)	Overflow(%)	F-score(%)	VSTG weight V
-	79.18	17.81	80.66	1

$E2$: Visual Concept STG; optimal threshold T (selected by exhaustive search on the test dataset)

	Visual concepts selected according to their AP				Visual concepts selected according to their ΔAP			
Concept Num. J	Coverage(%)	Overflow(%)	F-score(%)	Weight V	Coverage(%)	Overflow(%)	F-score(%)	Weight V
10	76.94	39.79	67.55	0	76.84	44.53	64.43	0
20	71.47	37.53	66.67	0	75.23	40.63	66.37	0
30	76.17	35.14	70.06	0	75.6	34.29	70.31	0
40	74.43	29.39	72.47	0	**72.56**	**24.35**	**74.07**	0
50	67.72	25.64	70.88	0	67.46	24.53	71.24	0
60	66.04	23.93	70.70	0	66.98	22.22	71.98	0
70	71.08	30.39	70.34	0	65.03	23.3	70.38	0
80	67.55	27.89	69.76	0	65.22	24.47	70.00	0
90	**74.02**	**26.43**	**73.79**	0	67.96	21.2	72.98	0
101	76.84	29.39	73.59	0	76.84	29.39	73.59	0

$E3$: Combination of Visual Concept STG and VSTG; optimal threshold T and VSTG weight V (both selected by exhaustive search on the test dataset)

	Visual concepts selected according to their AP				Visual concepts selected according to their ΔAP			
Concept Num. J	Coverage(%)	Overflow(%)	F-score(%)	Weight V	Coverage(%)	Overflow(%)	F-score(%)	Weight V
10	76.46	11.99	81.83	0.74	77.61	12.83	82.11	0.8
20	80.07	15.4	82.27	0.86	80	13.42	83.16	0.69
30	80.64	15.32	82.61	0.65	85.4	20.38	82.41	0.73
40	78.59	12.2	82.94	0.59	82.25	15.47	83.37	0.64
50	80.78	15.01	82.83	0.61	**80.69**	**13.66**	**83.42**	0.66
60	**79.95**	**13.21**	**83.23**	0.69	80.09	13.07	83.37	0.7
70	80.32	15.48	82.37	0.78	80.21	13.96	83.02	0.59
80	80.15	16.54	81.77	0.71	80.45	16.14	82.12	0.83
90	80.58	16.75	81.89	0.88	79.88	16.59	81.61	0.85
101	79.61	17.04	81.25	0.83	79.61	17.04	81.25	0.83

$E4$: Combination of Visual Concept STG and VSTG; VSTG weight V selected automatically using an out-of-testset video; optimal threshold T (selected by exhaustive search on the test dataset)

	Visual concepts selected according to their AP				Visual concepts selected according to their ΔAP			
Concept Num. J	Coverage(%)	Overflow(%)	F-score(%)	Weight V	Coverage(%)	Overflow(%)	F-score(%)	Weight V
10	79.44	16.48	81.43	0.69	79.38	15.56	81.83	0.73
20	82.17	18.54	81.81	0.8	80.02	14.87	82.50	0.84
30	80.65	16.31	82.14	0.84	81.06	16.3	82.36	0.72
40	80.61	16.11	82.22	0.71	**82.86**	**16.58**	**83.14**	0.61
50	80.64	16.15	82.21	0.46	81.06	15.62	82.69	0.75
60	**80.14**	**15.13**	**82.44**	0.61	80.95	15.44	82.72	0.63
70	80.04	15.61	82.16	0.75	80.78	15.31	82.69	0.65
80	80.03	16.52	81.72	0.73	80.18	15.86	82.11	0.68
90	79.45	16.02	81.65	0.77	79.88	16.59	81.61	0.85
101	79.38	17.24	81.03	0.87	79.38	17.24	81.03	0.87

$E5$: VSTG; threshold T selected automatically using an out-of-testset video

Concept Num. J	Coverage(%)	Overflow(%)	F-score(%)	VSTG weight V
-	74.75	16.81	78.74	1

$E6$: Combination of Visual Concept STG and VSTG; threshold T and VSTG weight V selected automatically using an out-of-testset video

	Visual concepts selected according to their AP				Visual concepts selected according to their ΔAP			
Concept Num. J	Coverage(%)	Overflow(%)	F-score(%)	Weight V	Coverage(%)	Overflow(%)	F-score(%)	Weight V
10	81.17	19.27	80.95	0.69	74.61	10.07	81.56	0.73
20	83.23	21.08	81.02	0.8	77.65	12.96	82.08	0.84
30	**79.74**	**15.76**	**81.93**	0.84	82.44	18.88	81.77	0.72
40	79.28	15.67	81.73	0.71	**81.08**	**15.72**	**82.65**	0.61
50	69.11	9.5	78.37	0.46	83.92	18.85	82.51	0.75
60	78.49	14.45	81.87	0.61	79.17	14.83	82.06	0.63
70	78.85	15.12	81.75	0.75	80.36	15.23	82.51	0.65
80	82.38	19.31	81.53	0.73	83.3	20.07	81.58	0.68
90	70.17	11.46	78.29	0.77	80.67	17.99	81.33	0.85
101	79.02	17.21	80.86	0.87	79.02	17.21	80.86	0.87