Spatiotemporal Semantic Video Segmentation

E. Galmar [*1], Th. Athanasiadis [†3], B.Huet [*2], Y. Avrithis [†4]

[*] *Département Multimédia, Eurécom, Sophia-Antipolis, France*
[1] galmar@eurecom.fr [3] huet@eurecom.fr

[†] *Image, Video & Multimedia Systems Laboratory, NTUA, Greece*
[2] thanos@image.ntua.gr [4] iavr@image.ntua.gr

Abstract—**In this paper, we propose a framework to extend semantic labeling of images to video shot sequences and achieve efficient and semantic-aware spatiotemporal video segmentation. This task faces two major challenges, namely the temporal variations within a video sequence which affect image segmentation and labeling, and the computational cost of region labeling. Guided by these limitations, we design a method where spatiotemporal segmentation and object labeling are coupled to achieve semantic annotation of video shots. An internal graph structure that describes both visual and semantic properties of image and video regions is adopted. The process of spatiotemporal semantic segmentation is subdivided in two stages: Firstly, the video shot is split into small block of frames. Spatiotemporal regions (volumes) are extracted and labeled individually within each block. Then, we iteratively merge consecutive blocks by a matching procedure which considers both semantic and visual properties. Results on real video sequences show the potential of our approach.**

I. Introduction

The development of video databases has impelled research for structuring multimedia content. Traditionally, low-level descriptions are provided by image and video segmentation techniques. The best segmentation is achieved by the human eye, performing simultaneously segmentation and recognition of the object thanks to a strong prior knowledge about the objects' structures. To generate similar high-level descriptions, a knowledge representation should be used in computer-based systems. One of the challenges is to map efficiently the low-level descriptions with the knowledge representation to improve both segmentation and interpretation of the scene.

We propose to associate spatiotemporal segmentation and semantic labeling techniques for joint segmentation and annotation of video shots. From one hand, semantic labeling brings information from a domain of knowledge and enables recognition of materials and concepts related to the objects. From the other hand, spatiotemporal segmentation decomposes a video shot into continuous volumes that are homogeneous with respect to a set of features. These extracted volumes represent an efficient medium to propagate semantic labels inside the shot.

Various approaches have been proposed for segmenting video shot into volumes. 3D approaches take as input the whole set of frames and give coherent volumes optimizing a global criterion [1], at the expense of an important computational cost. A few methods provide mid-level description of the volumes. In [2], volumes are modeled by a gaussian mixture model including color and position. Another example is given in [3], where volumes are considered as small moving linear patches. We have previously demonstrated that with a 2D+T (time) method [4] we can obtain a good trade-off between efficiency and accuracy of the extracted volumes. Recent progress has been also observed for scene interpretation and the labeling of image regions. In [5], an experimental platform is described for semantic region annotation. Integration of bottom-up and top-down approaches in [6] provides superior results in image segmentation and object detection. Region growing techniques have been adapted to group low-level regions using their semantic description instead of their visual features [7].

The integration of semantic information within the spatiotemporal grouping process sets two major challenges. Firstly, region labeling is obtained by computing visual features and match them to the database, which induces an important computational cost. Secondly, the relevance of the semantic description depends also on the accuracy of visual descriptors, whose extraction requires sufficient area of the volumes. These considerations suggest that use of semantic information during the early stages of the segmentation algorithm would be highly inefficient and ineffective if not misleading. Therefore, we add semantic information when the segmentation has produced a relatively small number of volumes. To this aim, we introduce a method to group semantically spatiotemporal regions within video shots.

The paper is organized as follows: In section II we give an overview of the strategy. Section III introduces the graph representation used for video shots. Section IV and V details the building steps of our approach: the labeling of temporal volumes and its propagation to the whole shot, respectively. Finally, results are illustrated in section VI and conclusions are drawn in section VII.

II. Overview of the strategy

The overall framework for the application is shown in fig.1. The considered video sequences are restricted to single shots, i.e. video data has been captured continuously from the camera and there are no cuts. Because of occlusion, shadowing, viewpoint change or camera motion, object material is prone to important spatial and temporal variations that makes maintaining an object as a unique volume difficult. To overcome the limits of the spatiotemporal stage, a video shot is decomposed into a sequence of smaller Block of Frames (BOF).

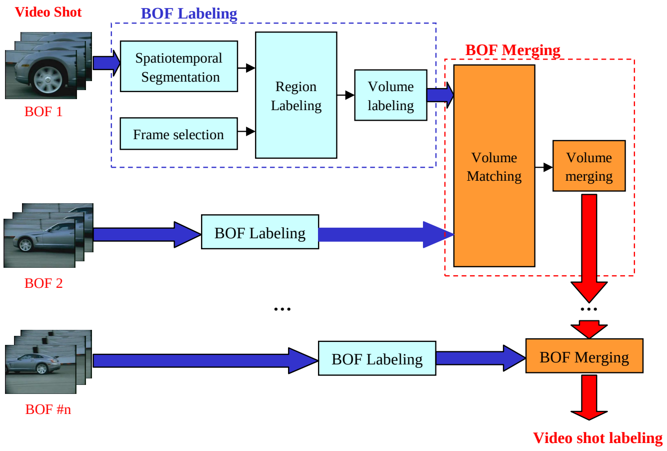


Fig. 1. The proposed framework for semantic video segmentation.

Semantic video shot segmentation is then achieved by an iterative procedure on the BOFs and operates in two steps, labeling of volumes within the BOF and merging with the previous BOF, which we will refer to as *intra-BOF* and *inter-BOF* processing respectively. During intra-BOF processing, spatiotemporal segmentation decomposes each BOF into a set of volumes. The resulting 2D+T segmentation map is sampled temporally to obtain several frame segmentation maps, each one consisting of a number of non overlapping regions. These regions are semantically labeled and the result is propagated within the volumes. A semantic region growing algorithm is further applied to group adjacent volumes with strong semantic similarity. During inter-BOF processing, we perform joint propagation and re-estimation of the semantic labels between consecutive video segments. The volumes within each BOF are matched by means of their semantic labels and visual features. This allows to extend the volumes through the whole sequence and not just within a short BOF. The semantic labels of the matched volumes are re-evaluated and changes are propagated within each segment. Finally both BOFs are merged and the process is repeated on the next BOF.

III. Graph Representation of Video Shots

Following MPEG-7 descriptions, one video shot is structured hierarchically in video segments. Firstly a shot is divided into M Blocks of Frames (BOF) B_i ($i \in [1, M]$), each one composed of successive frames F_t, $t \in [1, |B_i|]$. Spatiotemporal segmentation decomposes each B_i into a set of video regions (or volumes) S_{B_i}. Each volume $a \in S_{B_i}$ is subdivided temporally into frame regions $R_a(t)$, $F_t \in B_i$. Finally, frame segmentation at time t is defined as the union of frame regions of all volumes intersecting frame F_t: $S_t = \bigcup_{a \cap F_t \neq \emptyset} R_a(t)$. The elements composing the BOF are represented in fig.2.

A video segment (image or video shot) can represent a structured set of objects and is naturally described by an Attributed Relational Graph (ARG) [8]. Formally, an ARG is defined by spatiotemporal entities represented as a set of vertices V and binary spatiotemporal relationships represented as a set of edges E: $ARG \equiv \langle V, E \rangle$. Letting S_{B_i} be a segmentation of a BOF B_i, a volume $a \in S_{B_i}$ is represented in the graph by vertex $\mathrm{v}_a \in V$, where $\mathrm{v}_a \equiv \langle a, \mathcal{D}_a, \mathcal{L}_a \rangle$. $\mathcal{D}_a$ is a set of low-level MPEG-7 visual descriptors for volume a, while $\mathcal{L}_a$ is the fuzzy set of labels for that volume (defined over the crisp set of concepts C) with membership function μ_a:

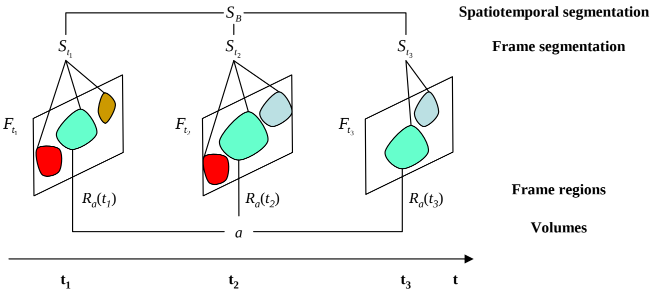


Fig. 2. Spatial and temporal decomposition of a BOF B_i.

$$\mathcal{L}_a = \sum_{i=1}^{|C|} c_i / \mu_a(c_i), \quad c_i \in C \tag{1}$$

Two neighbor volumes $a, b \in S_{B_i}$ are related by a graph edge $e_{ab} \equiv \langle (\mathrm{v}_a, \mathrm{v}_b), s_{ab}^{\mathcal{D}}, s_{ab}^{\mathcal{L}} \rangle$. $s_{ab}^{\mathcal{D}}$ is the visual similarity of volumes a and b, calculated from their set of MPEG-7 descriptors $\mathcal{D}_a$ and $\mathcal{D}_b$.. Several distance functions are used for each descriptor, so we normalize those distances linearly to the unit range and compute their visual similarity $s_{ab}^{\mathcal{D}}$ by their linear combination. $s_{ab}^{\mathcal{L}}$ is a semantic similarity value based on the fuzzy set of labels of the two volumes $\mathcal{L}_a$ and $\mathcal{L}_b$:

$$s_{ab}^{\mathcal{L}} = \sup_{c_i \in C} \left(t(\mathcal{L}_a, \mathcal{L}_b) \right), \quad a \in S, b \in N_a \tag{2}$$

where N_a is the set of neighbor volumes of a and t is a t-norm of two fuzzy sets. Intuitively, eq.2 states that the semantic similarity $s_{ab}^{\mathcal{L}}$ is the highest degree, implied by our knowledge, that volumes a and b share the same concept.

IV. Intra-BOF Labeling

To label a new BOF, we exploit the spatiotemporal segmentation to build visual and semantic description efficiently, using only a few frames. The following subsections present the criterion used for selecting these frames, the extraction of visual and semantic attributes of video regions and how those attributes are used for merging operations of volumes within the BOF.

A. Frame Selection

Once the segmentation masks are obtained for the whole BOF, region descriptor extraction and labeling tasks are substantially reduced by selecting a set of frames within the video segment. Choosing an important number of frames will lead to a complete description of the BOF but will require more time to process. On the contrary, using a single frame is more efficient but important volumes may not receive labels.

We consider a set of frames T and its corresponding frame segmentations $S_T = \{S_t\}$, $t \in T$ and measure the total span of the intersected volumes. Given a fixed size for T we choose the set T_{sel} that maximizes the span of the labeled volumes:

$$T_{sel} = \underset{T}{\operatorname{argmax}} \sum_{a \cap S_T \neq \emptyset} |a| \tag{3}$$

where $|a|$ is the size of volume a. Compared with fixed sampling, the criterion offers scalability for the extracted descriptors in function of the desired total volume span for the shot. Indeed the span increases with the number of frames selected.

B. Video Region Description

In previous work [5] we have shown how extracted visual descriptors can be matched to visual models of concepts. This region labeling process is applied to the selected frames (according to criteria discussed in section IV-A), resulting to an initial fuzzy labeling of regions with a set of concepts. The fuzzy set of labels $\mathcal{L}_a$ of a volume a is obtained by gathering the contributions from each frame region using a fuzzy aggregation operator :

$$\mu_a(c) = \frac{\sum_{t \in T_{sel}} \mathcal{A}(R_a(t)) \mu_{R_a(t)}(c)}{\sum_{t \in T_{sel}} \mathcal{A}(R_a(t))} \tag{4}$$

This operator weights the confidence degrees with the importance given to the frame regions. These weights $\mathcal{A}(R_a(t))$, are obtained by a measure of temporal consistency of frame regions.

Besides the semantic labeling, volumes are also described by low-level visual descriptors. Most MPEG-7 descriptors are originally intended for frame regions, but can be extended to volumes with the use of aggregation operators. For histogram-based descriptors, common operators are *mean*, *median* and *intersection* of bins. We select the *mean* operator since we consider homogeneous short-length volumes. In addition to descriptors, we also store the sizes and center of the volumes and its spatiotemporal bounding box for fast localization.

C. Semantic Volume Growing

Spatiotemporal segmentation usually creates more volumes than the actual number of objects present in the BOF. We examine how a variation of a traditional segmentation technique, the Recursive Shortest Spanning Tree (RSST) can be used to create more coherent volumes within a BOF. The idea is that neighbor volumes, sharing the same concepts, as expressed by the labels assigned to them, should be merged, since they define a single object.

To this aim, we modify the RSST algorithm to operate on the fuzzy sets of labels $\mathcal{L}$ of the volumes in a similar way as if it worked on low-level features (such as color, texture) [7]. The modification of the traditional algorithm to its semantic equivalent lies on the re-definition of the two criteria: (i) The similarity between two neighbor volumes a and b (vertices v_a and v_b in the graph), based on which graph's edges are sorted and (ii) the termination criterion. For the calculation of the semantic similarity between two vertices, we use $s_{ab}^{\mathcal{L}}$ defined in eq.2.

For one iteration of the semantic RSST, the process of volume merging decomposes in the following steps: Firstly, the edge e_{ab} that has the maximum semantic similarity $s_{ab}^{\mathcal{L}}$ is selected; vertices v_a and v_b are merged. Vertex v_b is removed completely from the ARG, whereas v_a is updated appropriately. This update procedure consists of two actions:

- Re-evaluation of the degrees of membership of the labels in a weighted average fashion from the union of the two volumes:

$$\mu_a(c) \leftarrow \frac{|a|\mu_a(c) + |b|\mu_b(c)}{|a| + |b|} \tag{5}$$

- Re-adjustment of the ARG edges by removing edge e_{ab} and re-evaluating the weights of the affected edges incident to a or b.

This procedure terminates when the edge e^* with maximum semantic similarity in the ARG is lower than a threshold, which is calculated in the beginning of the algorithm, based on the histogram of all semantic similarity values of the set of all edges E.

V. Inter-BOF Processing

In the previous section we dealt with segmentation and labeling of volumes within each single BOF. Here we examine how to extend volumes over consecutive BOF and for this purpose we develop techniques of visual and semantic volume matching. Semantic grouping is first performed on volumes with dominant concepts (i.e. concepts with high degree of confidence), then concepts are propagated temporally and spatially with the use of both semantic and visual similarity.

A. BOF Matching

We consider the merging of two successive BOF represented by their ARGs G_1 and G_2. It is not worth computing all volume matches between the two ARGs. As we consider continuous sequences, semantic objects are coherent spatially and temporally. In consequence, numerous matches can be pruned by exploiting spatiotemporal location of the volumes.

We establish temporal connections between G_1 and G_2 by selecting candidate matches from G_1 to G_2 and G_2 to G_1. Let G be the merged graph of G_1 and G_2. At the beginning, $G = G_1 \cup G_2$. Given vertices $\mathrm{v}_a \in G_1$ and $\mathrm{v}_b \in G_2$, v_a is connected to v_b in G if the bounding box of b intersects a truncated pyramid that represents the possible locations for a in the new BOF. The pyramid top base is defined by the bounding box of a. The bottom base is enlarged by a factor $D_s = v_{max} T_{max}$ where v_{max} is the maximum displacement between two frames and T_{max} is the height of the pyramid along the temporal axis. The connections are established in both forward and backward temporal directions. As a result, v_a owns an edge list of candidate matches $E_a = \{e_{ab} | \mathrm{v}_b \in G_2\}$. A list E_b is created similarly for v_b.

After creating the list of candidate matches, we match volumes with reliable or *dominant* concepts. A concept $c^* \in C$

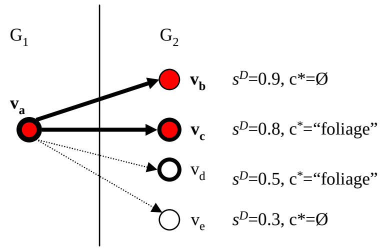


Fig. 3. Matching of dominant volumes. Dominant volumes are represented with thick circles.

is considered *dominant* for a volume $a \in G$ if the following condition is satisfied:

$$\left\{ \begin{array}{lll} \mu_a(c_1) & > & T_{dom} \\ \mu_a(c_1) & > & T_{sec}\mu_a(c_2) \end{array} \right. \tag{6}$$

c_1 and c_2 are respectively the concepts with highest and second highest degrees of membership. A dominant concept has degree of memberships above T_{dom} and is more important than all other concepts, with minimum ratio of T_{sec}.

The best match for one dominant volume may not be dominant because its visual appearance changes during the sequence. For this reason, we match either dominant volumes that have sufficient visual similarity or one dominant volume to any volume in case they have perfect visual match. The criterion to match a dominant volume a to a volume b, $e_{ab} \in E_a$, is based on both semantic and visual attributes. Let c_a^* and c_b^* be the dominant concepts of $\mathcal{L}_a$ and $\mathcal{L}_b$. If b is dominant but $c_a^* \neq c_b^*$, then no matching is done. In case c_b^* is empty, then e_{ab} has to be the best visual match from a, otherwise we compute the normalized rank of the visual similarity $s^{\mathcal{D}}$ in decreasing order, whose values do not depend of the descriptors used. Formally the criterion is validated if:

$$\left\{ \begin{array}{ll} rank\left(s_{ab}^{\mathcal{D}}\right) = 1 & \text{if} \quad c_b^* = \emptyset \\ \left\{ \begin{array}{l} c_a^* = c_b^* \\ \frac{|E_a| - rank\left(s_{ab}^{\mathcal{D}}\right)}{|E_a|-1} > T_s \end{array} \right. & \text{otherwise} \end{array} \right. \tag{7}$$

T_s indicates the tolerance allowed on visual attributes. When T_s is close to 1, only the best visual match is considered. If T_s is set to 0.5, half of the matches are kept.

The aforementioned procedure is illustrated fig.3. In the example, v_a is linked to v_c as it shares the same concept "foliage" and the visual similarity is the second best ($s_{ac}^{\mathcal{D}} = 0.8$). v_a is also linked to v_b since the similarity between a and b is the best one ($s_{ab}^{\mathcal{D}} = 0.9$). v_d is not matched even if it shares the same dominant concept as they are visually different from v_a. Indeed only dominant matches with good similarity are kept.

Since region and volume labeling are processes with a certain degree of uncertainty, reliable semantic concepts do not emerge from every volume, either due to the limited domain of the knowledge base, the imperfections of the segmentation, or the material itself. Therefore, we introduce volume matching using low-level visual attributes, expecting the semantics of these volumes to be recognized with more certainty in a subsequent part of the sequence. To avoid propagating matching errors and hamper the accuracy of the volumes, we only consider the matches with the strongest similarities and we are most confident in. Let e_a^* and e_b^* be the edges in lists E_a and E_b which have maximum visual similarity. a and b are matched and e_{ab} is a first best match, i.e. $e_{ab} \equiv e_a^* \equiv e_b^*$.

B. Update and Propagation of Labels

After the matching process, volumes are merged and their semantic and visual properties are computed using the aggregation operators, defined eq. 4. For this reason, new evidence for semantic similarity can be found in the merged graph as new dominant volumes are likely to be found. We do not merge further these volumes at this stage, so as to keep the accuracy of the visual description as they may correspond to different materials belonging to the same concept. Instead of this, the concepts of dominant volumes are propagated in the merged graph G. Let a be a non-dominant volume, $\mathrm{v}_a \in G$; we define a set of candidate dominant concepts $C_a = \{c \in C | \mu_a(c) > T_c\}$. For a concept $c \in C_a$, we compute the degrees of membership $\mu_a'(c)$ resulting from the aggregation of v_a and its neighbor vertices in G with dominant concept c:

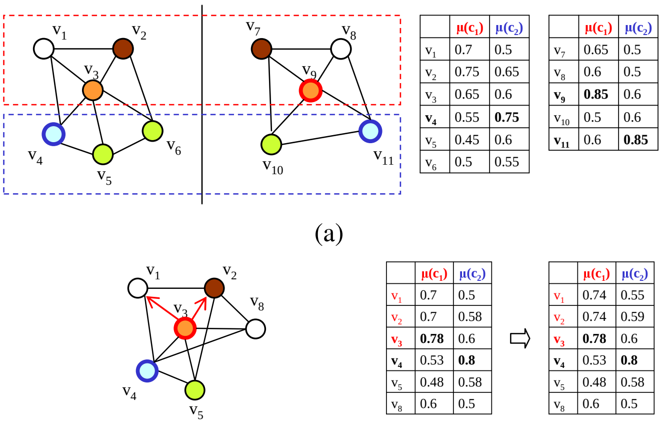
$$\mu_a'(c) = \frac{\sum_{b \in N_a^c} |b| \mu_b(c)}{\sum_{b \in N_a^c} |b|} \tag{8}$$

where $N_a^c = a \cup \{b \in N_a | c_b^* = c\}$ is the aforementioned neighborhood and $|b|$ is the current size of volume b. The concept $c^* \in C_a$, maximizing $\mu_a'(c)$, is selected and all degrees of membership of $\mathcal{L}_a$ and the size $|a|$ are updated by the aggregation of volumes in $N_a^{c^*}$. This propagation is performed in the whole graph G recursively. Let G^D be the subgraph of G containing only the dominant volumes of G and their incident edges. Once non-dominant volumes in G are processed, new dominant volumes may emerge in the subgraph $G' = G - G^D$. The update procedure is repeated considering G' as the whole graph until no more dominant volumes are found: $G^D = \emptyset$. Consequently, degrees of membership of non-dominant volumes tend to increase using the neighborhood context, correcting the values from the initial labeling.

Fig.4 gives an example of the inter-BOF merging and propagation of labels after that. The ideal semantic segmentation would be composed of two objects with dominant concepts c_1 and c_2. Before merging, a few dominant volumes are detected (v_4, v_9, v_{11}) in the two BOFs. After merging (fig.4(b)) the degrees of membership are re-evaluated according to eq. 5 and semantic weights are computed on the new edges. New evidence for semantic similarity is found between volumes (v_3, v_1) and (v_3, v_2), since v_3 has been matched with dominant volume v_9. Thus, due to propagation of concept c_1, v_1 and v_2 are linked to the dominant volume v_3 and their degrees of membership are increased according to eq. 8.

VI. Experimental Results

We illustrate the potential of the method on a set of examples. The knowledge domain encompasses various elements encountered in a natural scene, such as "sea", "sky", "foliage"



	$\mu(c_1)$	$\mu(c_2)$
v_1	0.7	0.5
v_2	0.75	0.65
v_3	0.65	0.6
$\mathbf{v_4}$	0.55	**0.75**
v_5	0.45	0.6
v_6	0.5	0.55

	$\mu(c_1)$	$\mu(c_2)$
v_7	0.65	0.5
v_8	0.6	0.5
$\mathbf{v_9}$	**0.85**	0.6
v_{10}	0.5	0.6
$\mathbf{v_{11}}$	0.6	**0.85**

	$\mu(c_1)$	$\mu(c_2)$
v_1	0.7	0.5
v_2	0.7	0.58
$\mathbf{v_3}$	**0.78**	0.6
$\mathbf{v_4}$	0.53	**0.8**
v_5	0.48	0.58
v_8	0.6	0.5

⇨

	$\mu(c_1)$	$\mu(c_2)$
v_1	0.74	0.55
v_2	0.74	0.59
$\mathbf{v_3}$	**0.78**	0.6
$\mathbf{v_4}$	0.53	**0.8**
v_5	0.48	0.58
v_8	0.6	0.5

Fig. 4. Merging of two BOFs. (a) Matching between two BOF. (b) Merging of a BOF and update of semantic labels. Ideal semantic segmentation is represented by the dashed boxes. Matched volumes are marked with similar colors, and dominant volumes are indicated with thick circles. Here, $T_{dom} = 0.75$ and $T_{sec} = 1.25$.

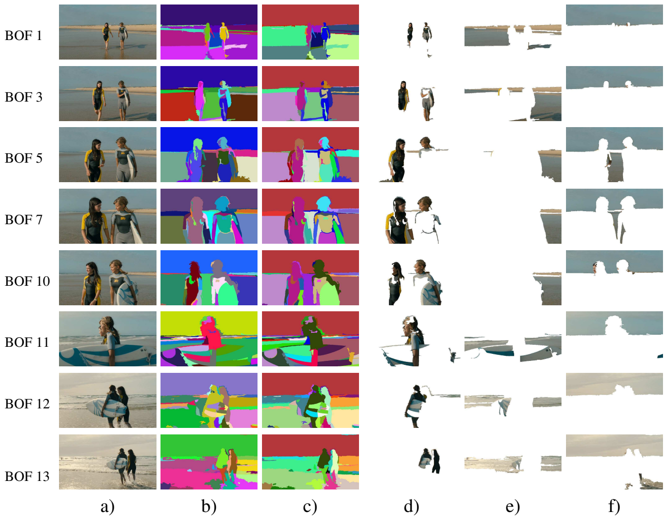


Fig. 5. Video semantic segmentation. (a) Frames in various BOF. (b) Spatiotemporal segmentation. (c) Semantic segmentation and inter-BOF matching. Volumes are extended throughout the shot (note the consistency in coloring). (d) Concept "person". (e) "sea". (f) "sky".

or "person". The proposed example sequences are composed of 650 and 100 frames, respectively. The BOF duration in the second sequence is $|B| = 10$ frames while for the first sequence we increase the duration to $|B| = 50$ frames, to show the behavior of the method at a larger scale while maintaining reduced computational costs.

The first example shows two girls walking on the beach (fig.5). Firstly, the girls are approaching the camera (BOF 1-5). Then they are observed in a close-up view (BOF 6-10). Finally the camera rotates quickly by 180 degrees to shoot them backside. Relevant concepts "person", "sky" and "sea" are detected within the shot. First we can see that the sky area is recognized all along the sequence. Although its aspect slightly changes at the end, it is still detected as dominant in the labeling stage and thus merged as a single volume. We can notice that isolated areas are also labeled "sky", as their material is visually close to this concept (BOF 5, 13). For the same reason, only part of the sea is identified at the right. In contrast, the left part is not dominant, but is correctly grouped by visual matching from BOF 3 to 10. After that, the sea areas are detected easily being shot in front view. The detection of "person" is more challenging since the related object includes different materials. In BOF 1 each silhouette is identified correctly standing as a single volume. The left girl's area is propagated from BOF 3 to 10. After that point it is completely occluded in BOF 11 and the concept is re-detected within a new volume in BOF 13. For the girl on the right the labeling is more uncertain as part of her suit and head have been confused with the background area (BOF 5, 7, 11). However, the upper part is still detected and propagated from BOF 5 to 9 and from 10 to 12 while the view is changing.

The second example shows a woman talking in front of her car (fig.6). The detected concepts include "person" and "foliage". The head and the coat both belong to the "person" concept and can be viewed as a single object, but are still separated in the semantic segmentation (fig.6(c)), which is an advantage as they are visually different. In BOF 4 only the coat is recognized (fig.6(d)). The reason is that the head has been partly confused with the background in the spatiotemporal segmentation. In such case, the volume is not matched, as its visual properties are different from the other volumes in the previous and subsequent BOF. In the right part of the sequence, the upper branches are well identified as "foliage" and are merged in a single volume from BOF 1 to 4 (fig.6(c)). From BOF 6 to 8, the branches are occluded by the woman. As a consequence the volumes are more fragmented and less homogeneous, so they are not linked to the previous part of the sequence. In BOF 10, the volume material in this area is homogeneous and the branches are correctly identified again.

		Foliage	**Person**	**Sea**	**Sky**	**Overall**
Ex.1	*Acc*	x	0.74	0.87	0.96	0.89
	Score	x	0.62	0.65	0.78	0.71
Ex.2	*Acc*	0.81	0.86	x	x	0.84
	Score	0.55	0.64	x	x	0.61

TABLE I
EVALUATION OF THE SEGMENTATION RESULTS.

Evaluation of the results for the above sequences is presented in table I. Each concept is associated to a semantic object (ground truth). The accuracy measure (*Acc*) [9] relates to the quality of the segmented volumes (fig.5-6(c)), unifying precision and recall. The evaluation score [7] gives a further measure of belief for the object labeling in every image. Unsurprisingly concept "sky" obtains the best result for all measures. For "foliage" sparse texture of the material and fragmentation of the volumes result in a lower score of 0.55. Concept "sea" has a higher detection score of 0.65, color and texture being relatively stable. Concept "person" is detected although some background can be included in the object

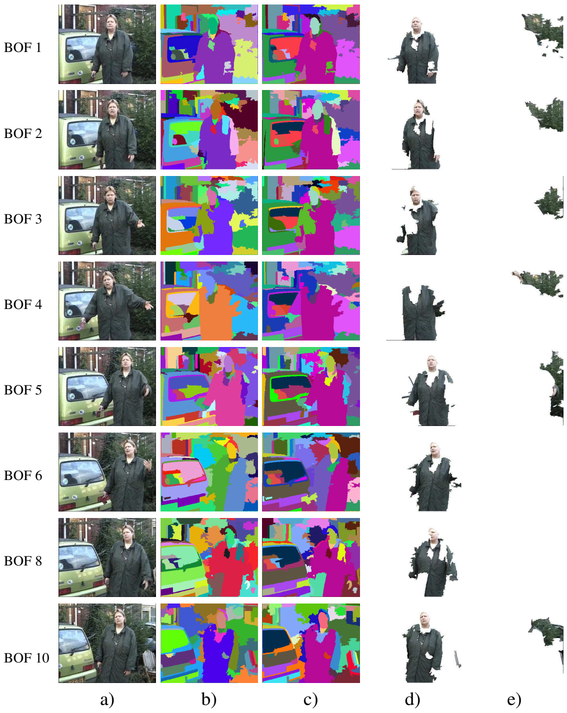


Fig. 6. Video semantic segmentation. (a) Frames in various BOF. (b) Spatiotemporal segmentation. (c) Semantic segmentation and inter-BOF matching. (d) Concept ”person”. (e) ”foliage”.

($Acc = 0.74$ for Ex.1). Finally, the overall detection scores are 0.71 for Ex.1 and 0.61 for Ex.2.

We further analyze the effects of the BOF decomposition on the efficiency of the approach. Fig.7 shows the repartition of the overall running time for the sequence of the first example (650 frames). The procedure is composed of four steps: (i) spatiotemporal segmentation, (ii) visual descriptor extraction and region labeling with the knowledg-assisted analysis system (KAA [5]), (iii) the construction of the ARGs (including the semantic RSST) and (iv) the inter-processing stage that merges the BOFs. Processing frames independently ($|B| = 1$) generates an important computational cost because of the labeling of every image of the sequence. The impact on the overall complexity is reduced with the spatiotemporal scheme ($|B| > 1$) that allows temporal sampling of the frames. For the evaluation, a single frame has been selected for each block, so that running time decreases inversely with the BOF length. Regarding the other components, we can notice that large BOF sizes lead to increase the time required for producing the spatiotemporal segmentation of the BOF. However, the additional cost is largely compensated with the gain in the region labeling stage. For the final merging stage, the running time for different BOF sizes is comparable. Indeed, the step is dominated by loading and updating the frame segmentation maps of which number does not depend of the BOF size, while the merging of the ARGs has lower complexity.

Overall, the gain with the proposed approach reaches a factor up to 12 ($|B| = 50$). Thus, the analysis shows the benefit of the framework in terms of complexity, extending single image annotation to continuous sequences efficiently.

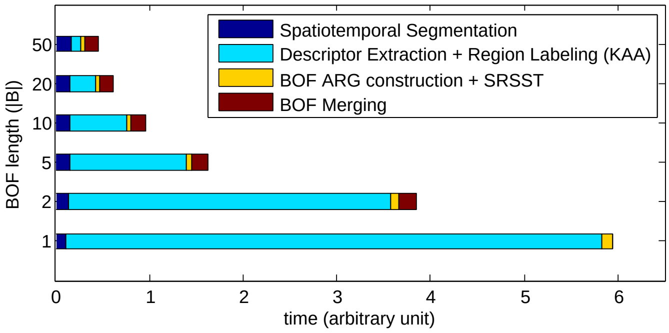


Fig. 7. Repartition of the overall running time of the first example, in function of the BOF length. Complexity is reduced when the BOF length increases.

VII. Conclusions

This paper presents a new approach for simultaneous segmentation and labeling of video sequences. Spatiotemporal segmentation is presented as an efficient solution to alleviate the cost of region labeling, compensating semantic with visual information when the former is missing. Our approach groups volumes with relevant concepts together while maintaining a spatiotemporal segmentation for the entire sequence. This enables the segmented volumes to be annotated at a subsequent point in the sequence. First experiments on real sequences show that the application is promising, though enhancements can still be achieved in the early spatiotemporal segmentation and labeling stage. Further challenge will be to consider structured objects instead of materials, leading towards scene interpretation and detection of complex events.

Acknowledgment

This research was supported by the European Commission under Contract FP6027026-K-Space.

References

[1] Y. Li, J. Sun, and H.-Y. Shum, “Video object cut and paste,” in *SIGGRAPH*, 2005.

[2] H. Greenspan, J. Goldberger, and A. Mayer, “Probabilistic space-time video modeling via piecewise gmm,” *IEEE Trans. PAMI*, vol. 26, no. 3, pp. 384–396, Mar. 2004.

[3] D. DeMenthon and D. Doermann, “Video retrieval using spatio-temporal descriptors,” in *ACM MM*, 2003, pp. 508–517.

[4] E. Galmar and B. Huet, “Graph-based spatio-temporal region extraction,” in *ICIAR*, 2006, pp. 236–247.

[5] T. Athanasiadis, V. Tzouvaras, V. Petridis, F. Precioso, Y. Avrithis, and Y. Kompatsiaris, “Using a multimedia ontology infrastructure for semantic annotation of multimedia content,” in *5th Int’l Workshop on Knowledge Markup and Semantic Annotation*, 2005.

[6] E. Borenstein, E. Sharon, and S. Ullman, “Combining top-down and bottom-up segmentation,” in *8th Conference on Computer Vision and Pattern Recognition Workshop*, 2004.

[7] T. Athanasiadis, P. Mylonas, Y. Avrithis, and S. Kollias, “Semantic image segmentation and object labeling,” *IEEE Trans. Circuits ans Systems for Video Technology*, vol. 17, March 2007.

[8] S. Berretti, A. D. Bimbo, and E. Vicario, “Efficient matching and indexing of graph models in content-based retrieval,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 12, pp. 1089–1105, Dec. 2001.

[9] F. Ge, S. Wang, and T. Liu, “Image-segmentation evaluation from the perspective of salient object extraction,” in *CVPR*, 2006, pp. 1146–1153.