A RESOURCE ALLOCATION FRAMEWORK FOR SUMMARIZING TEAM SPORT VIDEOS

Fan Chen and Christophe De Vleeschouwer
TELE, Université catholique de Louvain

ABSTRACT

We propose a flexible summarization framework for team-sport videos, which is able to integrate both the knowledge about displayed content (e.g. level of interest, type of view, etc.), and the individual (narrative) preferences of the user. Our framework builds on the partition of the original video sequence into independent segments, and create local stories by considering multiple ways to render each segment. We discuss how to segment videos based on production principles, and design the benefit function to evaluate various local stories from a segment. Summarization by selection of local stories is regarded as a resource allocation problem, and Lagrangian relaxation is performed to find the optimum. We use a soccer video to validate our framework in our experiments.

***Index Terms*—** Sport Video Summarization, Content Repurposing, Lagrangian Relaxation

1. INTRODUCTION

In many previous works related to sport video summarization [1][2], most of the efforts were put on automatic extraction of semantic information and production actions, while the summarization part is usually quite naive. In the present paper, we envision a novel user scenario of video summarization for production-room, as shown in Fig.1, where personalized video contents are produced based on meta-data provided directly from the production room. Limiting all required meta-data to several types of easily recordable production actions and annotated event data, we regard summarization as a constrained resource allocation problem by defining a set of candidate local stories for each video segment, and propose a novel framework to select and organize those local stories into sport video summaries. Compared with the previous work, our user-scenario has two major advantages. For content providers, it offers the capability to deploy additional services to achieve more profits, and it is labor efficient because all meta-data are effortlessly recordable from the production process. For customers, it offers personalization of access based on accurate and abundant meta-data, which should provide better user experiences.

2. SUMMARIZATION AS RESOURCE ALLOCATION

With all shot boundaries known, we divide a video into a sequence of clips. According to production principles of sport

This work is supported by the APIDIS project and Belgian NSF.

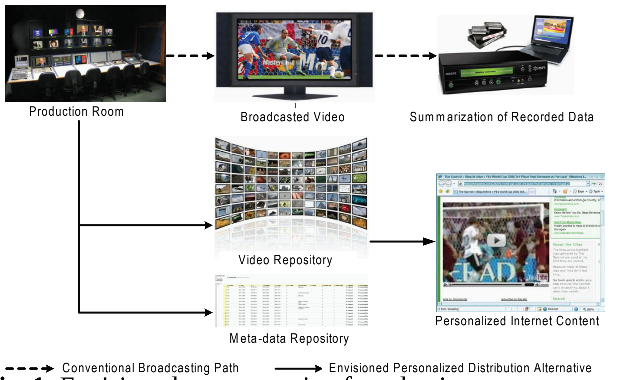

Fig. 1. Envisioned user scenario of production-room summarization. (Images are from the internet.)

videos[3], all non-replay clips in a sport video are not temporally overlapped, and no dramatic camera switching occurs during a critical action. Consequently, we envision personalized summarization in the divide and conquer paradigm, as shown in Fig.2. By analyzing patterns of camera switching, we organize these clips into non-overlapping segments. Each segment represents a short sub-story, which consists of successive clips that are semantically related. If we define a sub-summary as one way to organize clips within a segment, we regard the final summary as a collection of non-overlapping sub-summaries. Knowing view types of clips and events in the video, all potentially optimal combinations of clips within each segment are evaluated by their benefits and costs under specified user-preferences. We generate a universal set of candidate sub-summaries with various descriptive levels, and search for the best combination of sub-summaries which maximizes the overall benefit under user-preferred constraints.

Formally, assume that a video is divided into M segments, and the m-th segment $\mathbf{S}_m$ is a set of successive clips, which starts from time t_m^S and ends at t_m^E within the video. Collecting at most one sub-summary for each segment, we form the final story. If a sub-summary $\mathbf{s}_m$ is generated for the m-th segment, the overall benefit of produced summary is defined as accumulated benefits of all selected sub-summaries, i.e.,

$$\mathcal{B}(\{\mathbf{s}_m\}) = \sum_m \mathcal{B}(\mathbf{s}_m) \quad (1)$$

with $\mathcal{B}(\mathbf{s}_m)$ being defined in Section 4. The cost brought by using $\mathbf{s}_m$ is its length, which is defined as $|\mathbf{s}_m| = \sum_{i\in\mathbf{s}_m} |\tau_i^{\mathrm{E}} - \tau_i^{\mathrm{S}}|$. We use τ_i^{S} and τ_i^{E} to denote the starting time and ending time of the i-th clip in the broadcasted video. Our major task is to search for all optimal sub-summaries $\{\mathbf{s}_m\}$ that maximize the total payoff

$$\{\mathbf{s}_m^*\} = \arg\max_{\{\mathbf{s}_m\}} \mathcal{B}(\{\mathbf{s}_m\}) \quad (2)$$

under the length constraint $0 = \sum_m |\mathbf{s}_m| - u^{\mathrm{LEN}}$. u^{LEN} is the user-preferred length of summary. The above problem is equal to the well-known allocation problem of limited resources, if we modify the above constraint into a loose constraint of resource $\sum_m |\mathbf{s}_m| \leqslant u^{\mathrm{LEN}}$.

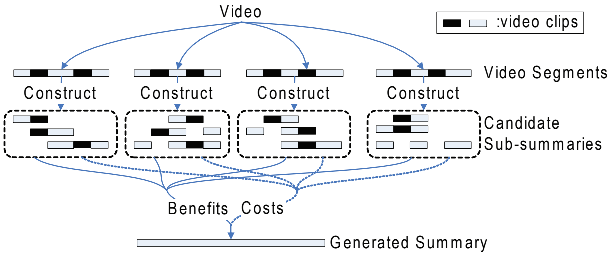

Fig. 2. Summarization in a divide-and-conquer paradigm.

A possible solution to solve the allocation problem of limited resources is to use the Lagrangian relaxation [4], whose main theorem reads that if λ is a non-negative Lagrangian multiplier and $\{\mathbf{s}_m^*\}$ is the optimal set which maximizes

$$\mathcal{L} = \sum_m \mathcal{B}(\mathbf{s}_m) - \lambda \sum_m |\mathbf{s}_m| \tag{3}$$

over all possible $\{\mathbf{s}_m\}$, then $\{\mathbf{s}_m^*\}$ maximizes $\sum_m \mathcal{B}(\mathbf{s}_m)$ over all $\{\mathbf{s}_m\}$ such that $\sum_m |\mathbf{s}_m| \leqslant \sum_m |\mathbf{s}_m^*|$. In our context, this theorem says that if $\{\mathbf{s}_m^*\}$ produced a summary with unconstrained maximum interest with a summary length being $\sum_m |\mathbf{s}_m^*|$, then $\{\mathbf{s}_m^*\}$ provides the best summary that could be produced without using more frames than $\sum_m |\mathbf{s}_m^*|$. From the curves of $\mathcal{B}(\mathbf{s}_m)$ with respect to their corresponding summary length $|\mathbf{s}_m|$, the collection of points with the same slope λ produces one unconstrained optimum. Different choices of λ lead to different summary lengths, and λ corresponding to a specified u^{LEN} could be achieved by trial and error procedures. If we construct a set of convex hulls from the curves of $\mathcal{B}(\mathbf{s}_m)$ with respect to $|\mathbf{s}_m|$, we can use a greedy algorithm to search for the optimum under the constraint u^{LEN}.

In Fig.3, a diagram summarizes the working flow of our framework. The rest of the paper is devoted to explain how to segment the video into segments, and how to derive benefit and cost for the multiple narrative options of each segment.

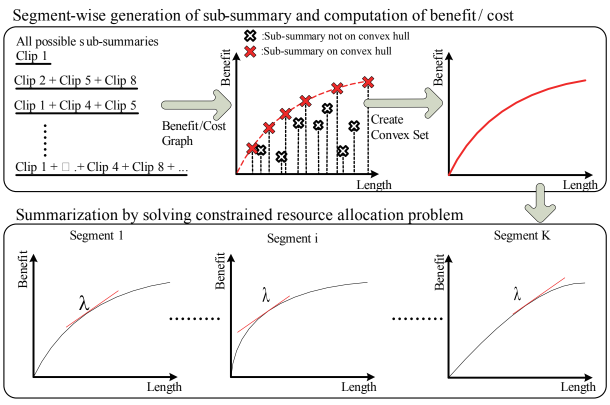

Fig. 3. Working flow in our summarization framework.

3. VIDEO SEGMENTATION

In Fig.4, we explain the segmentation rule that we envisioned for sport videos. It is worth noting that the proposed segmentation process only relies on information that are directly available from the production process. In other words, we do not assume any complex hand-made annotation process, or sophisticated automatic analysis of the video sequence to segment it. A general diagram of state transition in one round of offense/defense is given in the left of Fig.4. Starting with a kick-off type action, the offensive side makes trials of score after several passing actions. This trial ends up in one of three possible results: being intecepted, scoring, or being an opportunity. Before a new round, exceptional actions might happen, which include foul, medical assistance, and player exchange. We regard the state chain from the action-start to one of the results as a semantically complete segment, and separate exceptional actions in the state graph from the main action as individual segments.

State transition motivates scene switching, and is thus reflected in the production actions. We exploit this observation to segment the video based on the monitoring of production actions, instead of (complex) semantic scene analysis tools. In Fig.4, we show the typical view structure of a segment. One segment usually starts with a close view for highlighting the player who kicks off. A sequence of far view and medium view will be the major part of a segment to tell the story. After a key event is finished, some close-up shots might be given to raise the emotional level. According to the importance of the corresponding event, replay clips might also be appended. Note that close view, medium view and replay are all optional. Based on this structure, we divide the video into a series of segments. Although there are some complex cases, e.g., the offensive side tries many times of shooting, our segmentation rule is still applicable to them, because the producer will not switch the view type during those periods due to the tightness of match. A similar view structure was used in Ref.[1] to help detect exciting events(i.e., game parts with both close-up scenes and replays).

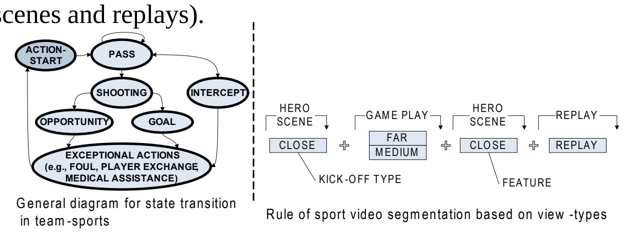

Fig. 4. Rule for segmentation of team-sports video.

4. CONSTRUCTION OF SUB-SUMMARIES

We assume that manual annotations are provided at production time to identify most interesting actions in terms of game relevance and emotional relevance. Such assumption is in accordance with the current practice in production rooms, as long as the annotations are only provided for the key actions of the game, in the form of a star rating system. Our approach follows this practical use case. Hence, each segment is assumed to be chararcterized by (a possibly empty) set of annotations, defining the game relevance and emotional level of the action(s) displayed within the segment[1]. In this section, we explain how the levels of interest defined by those annotations can be propagated to clips (Section 4.1), and translated in terms of benefit for a set of local stories considered for the segment (Section 4.2).

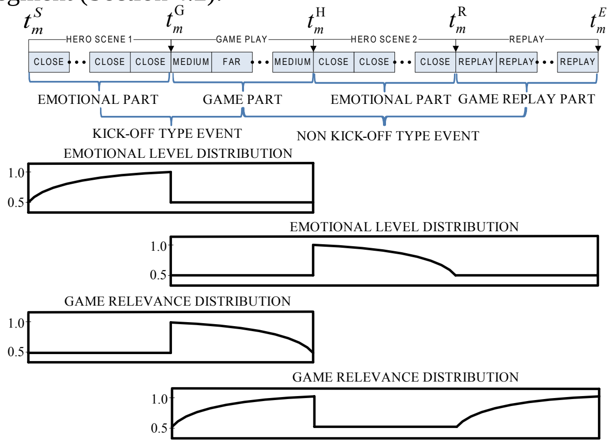

Fig. 5. Assigment of game relevance and emotional level.

4.1. Interest Level of Each Clip

Each segment is characterized by a (possibly empty) set of annotations, defining the game relevance and emotional level of actions displayed within the segment. Considering production principles, we design the distribution of game relevance and emotional level within a segment as shown in Fig.5.

For a kick-off type event which takes place in the beginning of the segment, we limit its influence within the first hero-scene and the game play part. Since the dominant player appears in the end of hero scene and commits this action in the begining of game play part, it is natural to let emotional level increase along with time evolving in the close view and let game relevance decrease in the game play. For non-kick-off game events, we design their distributions in Fig.5 based on the following facts: 1) Close views and replays are appended right after a critical action. A clip closer to the hero scene or replay has a higher game relevance. 2) The hero scene usually starts with the most highlighted player and then moves to other less important objects. Accordingly, the emotional level in the hero scene should decrease along with time evolving. 3) The replay part is designed to play the critical action in the same temporal order, which means that the game relevance is also increasing along with the evovling of replay. For other events, such as public events, player exchange and medical assitance, we assign uniform game relevance over its game part and uniform emotional level over its close-view part.

Mathmatically, assume that we have obtained N clips. For the i-th clip, we identify its scene type s_i, view type v_i, and replay status r_i. We set $s_i = 0$ for public scene and $s_i = 1$ for game scene, and set v_i to $0, 1, 2$ for the close, medium and far view, respectively. $r_i = 0$ for the i-th clip in the replay mode and $r_i = 1$ for normal game. Assume we have L_m events annotated for the m-th segment, $\mathcal{G}_l^G$ and $\mathcal{G}_l^E$ represent the 1 to 5 discrete grading of game relevance and emotional level assigned to the l-th event, respectively. The ratios of game relevance $\mathcal{R}_i^G$ and emotional level $\mathcal{R}_i^E$ of the i-th clip are computed by accumulating all related events, i.e.,

$$\mathcal{R}_i^E = \mathcal{T}_m^E \sum_l \mathcal{D}_{li}^E \mathcal{G}_l^E, \qquad \mathcal{R}_i^G = \mathcal{T}_m^G \sum_l \mathcal{D}_{li}^G \mathcal{G}_l^G \tag{4}$$

$\mathcal{T}_m^G$ is the length of game play in the segment, which includes all far/medium views, i.e., $\mathcal{T}_m^G = \sum_{i \in \mathbf{s}_m, v_i > 0} |\tau_i^{\mathrm{E}} - \tau_i^{\mathrm{S}}| s_i$. $\mathcal{T}_m^E$ is the length of emotional highlights, which consists of all non-replay close views, i.e., $\mathcal{T}_m^E = \sum_{i \in \mathbf{s}_m, v_i = 0} |\tau_i^{\mathrm{E}} - \tau_i^{\mathrm{S}}| s_i r_i$. $\mathcal{T}_m^G$ and $\mathcal{T}_m^E$ are introduced to avoid to favor short actions too much during the resource allocation process. The percentage of game relevance $\mathcal{D}_{li}^G$ and emotional level $\mathcal{D}_{li}^E$ of the l-th event on the i-th clip satisfy $1 = \sum_{i \in \mathbf{s}_m} \mathcal{D}_{li}^E = \sum_{i \in \mathbf{s}_m} \mathcal{D}_{li}^G$, and read

$$\mathcal{D}_{li}^E \propto \int_{\tau_i^{\mathrm{S}}}^{\tau_i^{\mathrm{E}}} 1 + \delta_{v_i,0} \tanh\left(\beta \frac{t - t_m^S}{t_m^G - t_m^S}\right) dt, \tag{5}$$

$$\mathcal{D}_{li}^G \propto \int_{\tau_i^{\mathrm{S}}}^{\tau_i^{\mathrm{E}}} 1 + (1 - \delta_{v_i,0}) \tanh\left(\beta \frac{t_m^H - t}{t_m^H - t_m^G}\right) dt \tag{6}$$

for kick-off events, and

$$\mathcal{D}_{li}^E \propto \int_{\tau_i^{\mathrm{S}}}^{\tau_i^{\mathrm{E}}} 1 + \delta_{v_i,0} \tanh\left(\beta \frac{t_m^R - t}{t_m^R - t_m^H}\right) dt, \tag{7}$$

$$\begin{aligned} \mathcal{D}_{li}^G \propto \int_{\tau_i^{\mathrm{S}}}^{\tau_i^{\mathrm{E}}} & 1 + (1 - \delta_{v_i,0})\delta_{r_i,1} \tanh\left(\beta \frac{t - t_m^G}{t_m^H - t_m^G}\right) \\ & + (1 - \delta_{v_i,0})\delta_{r_i,0} \tanh\left(\beta \frac{t - t_m^R}{t_m^E - t_m^R}\right) dt \end{aligned} \tag{8}$$

for non-kick-off events, where t_m^G, t_m^H, t_m^R are starting times of game play part, hero scene part, and replay part of the m-th segment, respectively, as shown in Fig.5. β is a parameter introduced to control the decaying speed and $\delta_{x,y}$ is the Kronecker delta. We use hyperbolic tangent function $\tanh(x)$ to model the decaying process, because it is bounded and is integrable which simplifies the computation of $\mathcal{D}_{li}^E$ and $\mathcal{D}_{li}^G$.

The interest level of the i-th clip is computed as

$$\mathcal{I}_i = \alpha \mathcal{R}_i^E + (1 - \alpha) \mathcal{R}_i^G; \tag{9}$$

α is a hyper-parameter of user preference to control the relative importance of emotional level and game relevance.

4.2. Benefits of Sub-summaries

We now explain the construction of sub-summary by clip selection within a segment. We use $a_{mi} = 1$ to represent the adoption status of selecting the i-th clip into the sub-summary $\mathbf{s}_m$, and $a_{mi} = 0$ for not selecting the clip. We propose to define the benefit gained by using $\mathbf{s}_m$ to represent the m-th segment as follows

$$\mathcal{B}(\mathbf{s}_m) = \sum_{i \in \mathbf{s}_m} \mathcal{I}_i \left(1 + \frac{a_{m(i-1)} + a_{m(i+1)}}{10}\right) \mathcal{P}(\mathbf{s}_m). \tag{10}$$

[1]In absence of annotation, a segment simply receives a reference unitary level of interest, both regarding game relevance and emotional interest.

The term of $a_{m(i-1)} + a_{m(i+1)}$ is included to add extra benefit from smoothness of story-telling by selecting successive two clips. $\mathcal{P}(\mathbf{s}_m)$ represents the penalty brought by redundancy in the sub-summary and other forbidden cases, e.g., the replay is selected while no normal play part is selected, i.e.,

$$\mathcal{P}(\mathbf{s}_m) = \left[\frac{\sum_{i \in \mathbf{s}_m, v_i > 0, r_i = 1} |\tau_i^{\mathrm{E}} - \tau_i^{\mathrm{S}}|}{2 \sum_{i \in \mathbf{s}_m} |\tau_i^{\mathrm{E}} - \tau_i^{\mathrm{S}}|} + \frac{1}{2} \right]^{\gamma} \mathcal{P}_2(\mathbf{s}_m). \quad (11)$$

where the term in the bracket shows the rate of redundancy brought by replays and close views. Higher γ tolerates less redundancy, which produces a summary with more but shorter sub-summaries with less replays, while lower γ produces a summary including less but longer sub-summaries with detailed replays. $\mathcal{P}_2(\mathbf{s}_m)$ switches its value between 0.1 and 1, according to whether $\mathbf{s}_m$ is of a forbidden form or not.

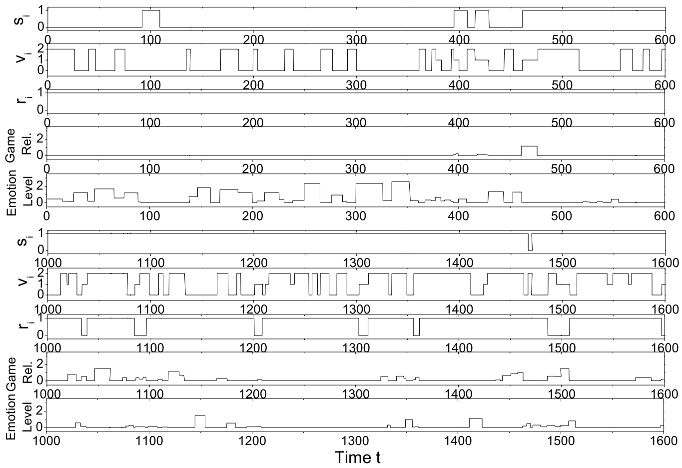

Fig. 6. Profiles of two selected periods and their game relevance and emotional levels.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

We use a soccer video for a whole game as experimental data to verify our summarization method. Along with the video, we received a list of 200 annotated events. We detect shot-boundaries based on the method in Ref.[5], replay by finding all replay-specified logos, and view-types by using the method in Ref.[2] from the video. The video is divided into 1106 clips after processing, and all experiments were applied to the whole video. Due to the limitation of page width, we only show quantitative data on two selected periods of the video , i.e., one from 0s to 600s to represent public events, and another from 1000s to 1600s to represent the game. Profiles of two selected periods and their game relevance and emotional levels are shown in Fig.6. Public events show higher emotional levels and lower game relevance than game events. A long event from 1425s to 1525s reflects the way we assign emotional level and game relevance in Fig.5.

By varying α and u^{LEN}, we plot selection status of clips in generated summaries in Fig.7. We confirm that 1) α can trade-off the summary between higher emotional level and closer game relevance. 2) No standalone replay exists as we designed in the framework. 3) To generate a longer summary, local stories expand in an adaptive way according to their view structures, rather than simply expanding their widths. Reviewers are invited to download demo videos in our website to have subjective evaluation of the results [6].

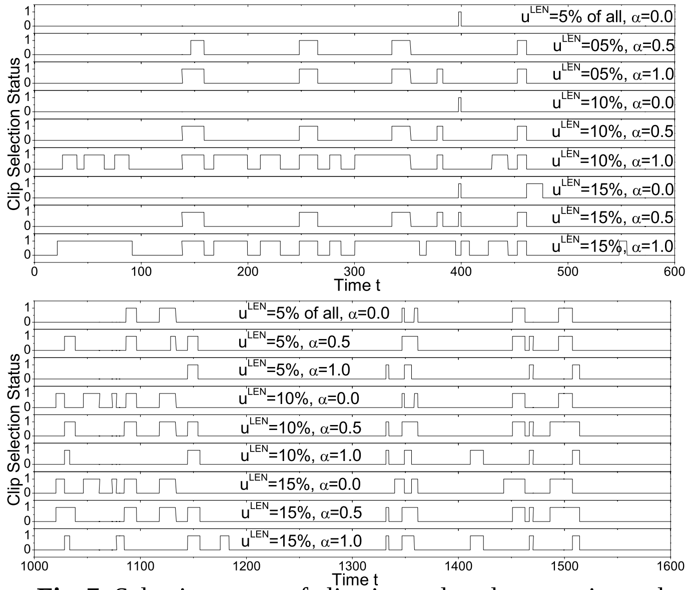

Fig. 7. Selection status of clips in produced summaries under different values of user-preference u^{LEN} and α.

For a long segment with many clips, we first compute convex sets for its hero scene, game play and replay separately, and then combine those results to produce the convex set for the whole segment. Our summarization method is real-time, which needs less than one second to summarize the video.

6. CONCLUSIONS

We proposed a framework for producing personalized summarization of sport videos. The approach is computationally efficient by taking divide-and-conquer strategy. The video is divided into many short clips. Within each local time segment, we deal with short clips carefully to organize the story-telling. Globally, we use Lagrangian relaxation to find the optimal selection of clips to form the final summary. Our method is also flexible in the sense that it supports different definition of benefit to customize the summarization process.

7. REFERENCES

[1] Li B., Pan H., and Sezan I. "A general framework for sports video summarization with its application to soccer," *ICASSP'03*, vol.**3** pp.169-172, 2003.

[2] Ekin A., Tekalp A.M., and Mehrotra R., "Automatic soccer video analysis and summarization," *IEEE Trans. Image Processing*, vol.**12** pp.796-807, 2003.

[3] Owens J., "Television sports production," *Focal Press*, 2007.

[4] Everett H., "Generalized lagrange multiplier method for solving problems of optimum Allocation of Resources," *Operations Research*, vol.**11** pp.399-417, 1963.

[5] Delannay D., Roover C., and Macq B., "Temporal alignment of video sequences for watermarking systems,"*Proc. SPIE*, vol.5020, pp.481-492, 2003.

[6] Demo Link: http://thetis.tele.ucl.ac.be/Apidis/chen/