Resource Allocation for Personalized Video Summarization

Fan Chen, Christophe De Vleeschouwer and Andrea Cavallaro

***Abstract*—We propose a hybrid personalized summarization framework that combines adaptive fast-forwarding and content truncation to generate comfortable and compact video summaries. We formulate video summarization as a discrete optimization problem, where the optimal summary is determined by adopting Lagrangian relaxation and convex-hull approximation to solve a resource allocation problem. To trade-off playback speed and perceptual comfort we consider information associated to the still content of the scene, which is essential to evaluate the relevance of a video, and information associated to the scene activity, which is more relevant for visual comfort. We perform clip-level fast-forwarding by selecting the playback speeds from discrete options, which naturally include content truncation as special case with infinite playback speed. We demonstrate the proposed summarization framework in two use cases, namely summarization of broadcasted soccer videos and surveillance videos. Objective and subjective experiments are performed to demonstrate the relevance and efficiency of the proposed method.**

***Index Terms*—Personalized Video Summarization, Resource Allocation, Adaptive Fast-Forwarding**

I. Introduction

Video summarization techniques are relevant for various applications, such as TV program/movie production, surveillance and e-learning [1] and may address different purposes, including fast browsing [2], information retrieval [3][4], behaviour analysis [5] and entertainment [6]. In order to generate from the source video(s) a well-organized and concise version that best satisfies the interest of a user, the most important requirement of summarization is comprehensibility. Other important criteria to judge summarization quality are personalization, visual comfort and the quality of story-telling. Personalization is essential for satisfying various user tasks and narrative preferences. Visual comfort increases when the flickering caused by abrupt scene transitions is reduced. The quality of story-telling depends on the integrity of the story with the inclusion of the most significant moments and the continuity of the summaries. We encapsulate these requirements into three properties: *completeness* (evaluating the amount of clearly presented events of interest in the summary), *comfort* (which decreases in the presence of flickering and abrupt story transitions) and *effectiveness* of time allocation (the relevance of the playback time assignment).

Video browsing can be seen as an information communication process between the video producer and the audience. Two kinds of information need to be considered for producing semantically relevant and enjoyable summaries, namely information associated to the still content of the scene and information associated to the scene activity. Information associated to the *still content* of the scene helps evaluating the importance of frames for producing semantically relevant summaries. Information associated to the *scene activity* is associated to the visual stimulus offered to the audience. An audience will get bored with a video with few stimuli (e.g., a long surveillance video without events), and will be overstressed with a video with an amount of stimuli beyond his visual comfort limits. This information is thus important in determining the attraction and enjoyment of summaries.

Conventional *content-truncation-based methods*, such as presenting a sequence of key frames or a sequence of moving images (video skimming), mainly maximize the transferred information associated to the still content during a constrained browsing period (e.g., using fast-browsing of highlights [7][8]). However, information extracted from still contents cannot model complex story-telling with strong dependency in its contents when the summary is presented as a video. As for visual comfort, an attractive and entertaining video content cannot be produced by simply collecting the most significant key frames. The amount of stimuli during continuous browsing of naive key frames would be too large due to significant frame differences [7]. Video skimming provides more visually comfortable results by reducing the amount of stimuli at the cost of sacrificing the information associated to less relevant events.

Conventional *fast-forwarding-based methods* mainly subsample frames based on information associated to scene activity, defined via optical flow [9] or the histogram of pixel differences [10]. By only evaluating changes in the scene, it is difficult to assure the semantic relevance of the summary. The application of pure fast-forwarding-based methods is also constrained by the fact the highest tolerable playback speed is bounded [11]. According to the observations in visual perception, attentional processes (defined by target selectivity, e.g., identifying a suspicious person in surveillance) usually

 This work was supported in part by the Japan Grant-in-Aid for Young Scientists (B)(No.23700110), by the Belgian NSF, by the Walloon Region project SPORTIC, and by the UK Engineering and Physical Sciences Research Council (EPSRC), under grant EP/K007491/1.

F. Chen is with the School of Information Science, Japan Advanced Institute of Science and Technology, Nomi 923-1211, Japan (e-mail: chenfan@jaist.ac.jp).

C. De Vleeschouwer is with the ICTEAM of Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium (e-mail: christophe.devleeschouwer@uclouvain.be).

A. Cavallaro is with the Centre for Intelligent Sensing, Queen Mary University of London, London E1 4NS, UK. (e-mail: andrea.cavallaro@eecs.qmul.ac.uk).

TABLE I
COMPARISON OF STATE-OF-THE-ART METHODS FOR VIDEO SUMMARIZATION

Content Presentation	*Task*	*Principles*	*Approach*	*Ref.*	*Content Type*
Key-frame extraction	Fast-browsing	Feature-based Indexing	Unsupervised Clustering	[13]	Generic
			Fuzzy Clustering	[2]	
			Discriminative Analysis	[14]	
		Semantic-based Indexing	Attention-based	[15]	Generic
		Sub-sampling for Min. Distortion	MINMAX Optimization	[8][16]	Generic
	Content-retrieval	Feature-based Indexing	Compressed Domain Retrieval	[17]	Generic
		Object-based Indexing	Spatial/Temporal Sampling	[3]	Surveillance
		Semantic-based Indexing	Multi-modal Meta-data	[18]	Soccer
	Behaviour analysis	Object-based Indexing	Action Key Poses	[18]	Human Action
Video skimming	Fast-browsing	Fixed Skimmed Length	Motion Attention	[19]	Generic
		Rule-based Creation	Audio/Video Skimming	[20]	Generic
		Filtering-based Creation	Priority Curve Algorithm	[21]	Soccer
		Story-based Creation	Semantic Relation Graph	[22]	Generic
	Content-retrieval	Feature-based Indexing	Shot-classification	[23]	Broadcasted News
		Object-based Indexing	Region Association in a Shot	[24]	Generic
		Semantic-based Indexing	Lexicon-driven Retrieval	[25]	Generic
	Behaviour analysis	Pattern Classification	Tactic Analysis	[5]	Soccer
	Video Enjoyment	Temporal Alignment	Audio/Video Synchronization	[6]	Soccer
		Rule-based Creation	Cinematic Rules	[26]	Soccer
Fast-forwarding	Fast-browsing	Feature-based Sub-sampling	Optical-flow-based	[9]	Surveillance
			Information-theory-based	[10]	
		Rule-based Sub-sampling	Smart-player	[27]	Generic
		Sub-sampling for Min. Distortion	Key-frame-based	[28]	Generic
	Content-retrieval	Sub-sampling for Max. Similarity	Generative-model-based	[29]	Generic
Video condensation	Browsing/Retrieval	Optimal Space Allocation	Ribbon Carving	[30]	Surveillance
			Video Synopsis	[31]	
			Online Video Condensation	[32]	
Video skimming and fast-forwarding	**Browsing/Enjoyment**	**Optimal Allocation of Playback Time**	**Resource Allocation**	**Proposed Method**	**Surveillance /Team-sport**

have even lower limits than non-attentional processes (without target selectivity) [12]. Under the request of a highly compact summarization, less relevant contents will need to be presented with too high playback speeds thus producing annoying visual artifacts, such as flickering [9] [10].

To overcome these limitations, we propose an approach that truncates contents with intolerable playback speeds and saves time resources for better rendering the remaining contents. We design a hybrid summarization method combining content truncation and adaptive fast-forwarding to provide continuous and complete summaries with improved visual comfort. Moreover, we provide a new perspective in understanding the motivations behind truncation-based and fast-forwarding-based summarization techniques. We select playback speeds from a set of discrete options, and introduce a hierarchical summarization framework to find the optimal allocation of time resources into the summary, which performs nonlinear computation of overall information in the summary and enables various story-telling patterns for flexible personalized video summarization. Other contributions include subjective observations on suitable playback speeds and a method for hot-spot detection by automatic extraction of group interactions from surveillance videos.

The paper is organized as follows. After a brief review of previous video summarization methods in Section II, in Section III we discuss a criterion that trades-off fast-forwarding and visual comfort. In Section IV we introduce the proposed summarization framework, along with the optimization techniques for global story organization. Section V discusses the application of the summarization framework to two use cases. Finally, we present experimental results in Section VI whereas Section VII concludes the paper.

II. RELATED WORK

We classify video summarization methods in three categories, based on their content presentation techniques: reorganization of story-telling, video condensation and adaptive fast-forwarding.

Reorganization of story-telling truncates less relevant content or changes its temporal order. Most methods based on key-frame extraction and video skimming belong to this category [8][13]. Early techniques extract a short video sequence of a desired length to maximize the included information, which results in minimizing the loss due to the skipped frames and/or segments. These methods generally differ in the definition of the similarity between the summary and the original video, and in the techniques used to maximize this similarity. They include methods to cluster similar frames/shots into key frames [2][7], and methods for constrained optimization of objective functions [8][16]. Other methods measure precision and recall rates of different events in soccer based on cinematic rules [26] or sound analysis [33]. Fast-forwarding methods that perform conventional key-frame extraction by minimizing the reconstruction error also belong to this category [28]. Since they attempt to preserve the initial content as much as possible, these methods are well suited to support efficient browsing. The motivation of end-users in viewing summaries is not limited to fast browsing of all clips in the whole video content. A summary can also be organized to provide information to special users, such as helping the coach to analyse the behaviour of players from their trajectories [5]. Summarization is also used for organizing music soccer sport videos, based

on the synchronization between video and music content [6]. Continuity of clips is important for story-telling [21]. Story organization is also considered via a graph model for managing semantic relations among concept entities [22]. For broadcasted soccer videos, personalized story-telling can be organized by assigning event significance [34] and extracting specified view types [35]. Summarization framework exists for enhanced personalization of story-telling to satisfy both narrative (including continuity, redundancy and prohibited story organization) and semantic audience preferences (i.e. favourite events/objects) [36][37]. Personalized summarization is also implemented as a "query" function to extract objects/events preferred by the user, via textual descriptors and, optionally, with interaction [3][4].

Video condensation considers the efficient rendering of object activities in summaries by embedding sequences of video objects into the seams of the video. A ribbon-carving-based method considers just the moving objects (rather than a whole frame) as the smallest processing unit [30]. Moving objects are first isolated from the videos and put into an object database. According to the requirements of the users, interesting objects are picked up from the database [31] or created on-line [32], and their activities are rendered in a synopsis video. However, video condensation fails to preserve the temporal order and relationship of multi-object activities.

Adaptive fast-forwarding condenses the video by adjusting the playback speeds. An intuitive consideration in adaptive fast-forwarding is to equalize the motion complexity in terms of optical flow in the summaries [9]. A fast-forwarding method based on the normalized intensity of motion vectors was also considered along with user specified target playback speeds [27]. However, motion vectors are not always consistent with scene complexity because of different zoom factors and because of the noise generated in the motion estimation phase. Summarization can also be interpreted as a query process, where the playback speed is adjusted according to the similarity between the frame and the target content [29]. Adaptive fast-forwarding can be considered from the perspective of information theory, with the goal of equalizing the scene complexity, represented by the statistical distance (alpha-divergence) between the frame difference and the learnt noise model [10]. Various visualization techniques for fast-forwarded summaries can be used [38]. Pure fast-forwarding is not suitable for highly compact summarization (e.g., a 10-minute summary of a 24-hour surveillance video), due to the maximum tolerable playback speed upper-bounded by the limitations of both visual perception and memory [11].

Table I presents a summary of related works based on content presentation, task, approaches and target content types.

III. A Criterion to Balance Playback Speed and Visual Comfort

In this section, we discuss video summarization and derive the corresponding mathematical criterion that enables us to balance playback speed and visual comfort.

Let a source video $\mathbf{V}$ be a sequence of N frames evenly sampled, $\mathbf{V} = \{I_t | t = 1, \cdots, N\}$. I_t represents the image data in the t^{th} frame. Given u^{L} as the user-specified constraint on the summary length, we formulate the summarization process as finding a sequence $\hat{\mathbf{V}} = \{(I_t, s_t) | t = 1, \cdots, N; s_t \in [0 \quad 1]\}$ subject to $\sum_{t=1}^{N} s_t = u^{\text{L}}$. s_t is the adjusted temporal distance (i.e. the inverse of playback speed) between the $(t-1)^{\text{th}}$ and t^{th} frames, and is normalized by the unit sample interval in the source video. $s_t = 0$ stands for infinite playback speed, which is equivalent to content truncation.

Conventional content truncation only allows to take s_t from $\{0, 1\}$, and searches for a subset of frames that maximizes the overall information

$$\hat{\mathbf{V}}^*_{\text{C}} = \arg\max_{\hat{\mathbf{V}}} \sum_{t=1}^{N} s_t f_t, \quad \text{s.t.} \sum_{t=1}^{N} s_t = u^{\text{L}}, \tag{1}$$

where f_t is the information associated to the still content in the t^{th} frame.

Adaptive fast-forwarding allows real values of s_t from $[0 \quad 1]$. Let a_t be the information associated to scene activity in the t^{th} frame of the source video. Adaptive fast-forwarding finds s_t that makes the adjusted information a_t/s_t proportional to the pre-specified target strength with the highest comfort C^{c}_t [9][10]:

$$\forall t, \frac{a_t}{s_t} \propto C^{\text{c}}_t \Rightarrow s_t \propto \frac{a_t}{C^{\text{c}}_t}. \tag{2}$$

This maximizes the visual comfort during video browsing in terms of C^{c}_t, and is computationally equivalent to

$$\hat{\mathbf{V}}^*_{\text{F}} = \arg\max_{\hat{\mathbf{V}}} \sum_{t=1}^{N} \left[\frac{a_t}{C^{\text{c}}_t}\right]^{1/2} [s_t]^{1/2}, \quad \text{s.t.} \sum_{t=1}^{N} s_t = u^{\text{L}}, \tag{3}$$

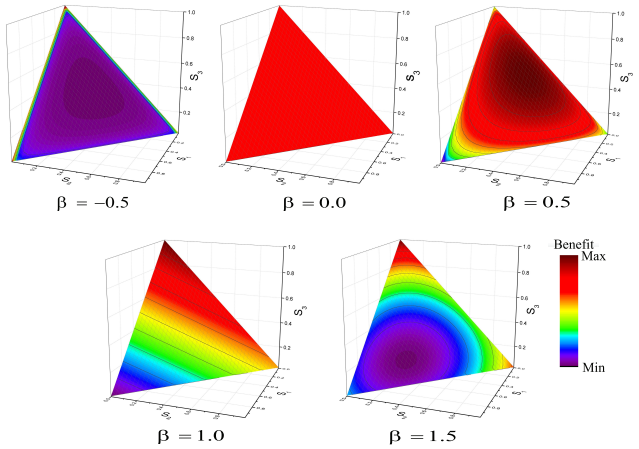
where $1/2$ assures linear proportionality in Eq.2.

The criterion in Eq.1 only considers the information associated to still content and fails to handle the redundancy in duplicated content and does not consider visual comfort. The criterion in Eq.3 considers the information associated to scene activity, which is not always consistent with the semantic relevance of the summary. Hence, it is necessary to include both types of information to produce semantically relevant and comfortable summaries. We therefore propose the following unified criterion

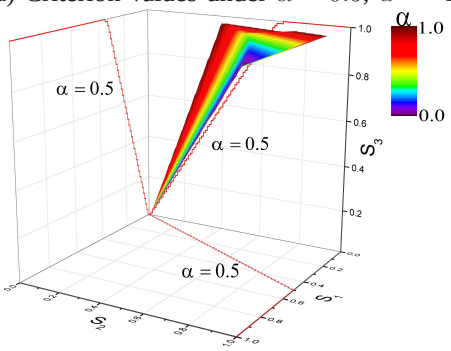
$$\hat{\mathbf{V}}^* = \arg\max_{\hat{\mathbf{V}}} \sum_{t=1}^{N} [f_t]^{\alpha_1} \left[\frac{a_t}{C^{\text{c}}_t}\right]^{\alpha_2} s_t^{\beta}, \quad \text{s.t.} \sum_{t=1}^{N} s_t = u^{\text{L}}. \tag{4}$$

Note that both Eq.1 and Eq.3 are abbreviated special cases of the above criterion.

Without loss of generality, let $\alpha_1 = \alpha_2 = \alpha$ to simplify the discussion. For simplicity of notation, let $\theta_t = f_t \frac{a_t}{C^{\text{c}}_t}$. In Fig. 1(a), we show the behaviour of the above criterion under $u^{\text{L}} = 1, \alpha = 0.5$ and various β, in an example case of $N = 3$ frames with different information values $\theta_1 = 1, \theta_2 = 2, \theta_3 = 3$ to investigate the distribution of s_t. For $\beta < 0$, it reaches infinity when s_t approaches 0, which takes place when contents are truncated, and becomes constant at $\beta = 0$, which makes the summarization problem irrelevant. A longer s_t will be assigned to frames with higher information when $0 < \beta$. When $\beta \geq 1$, it forms a convex function. In this case, it will simply assign $s_t = 1$ to frames in the descending



(a) Criterion values under $\alpha = 0.5$, $u^{\mathrm{L}} = 1$



(b) Optimal solutions of $\{s_t\}$ under $\beta = 0.5$, $u^{\mathrm{L}} \in (0, 3]$

Fig. 1. The behaviour of our criterion in a simple case of three frames $\theta_1 = 1, \theta_2 = 2, \theta_3 = 3$. a) a balanced distribution of playback time is achievable under $0 < \beta < 1$. The rectangle plane is for $\sum_{t=1}^{3} s_t = u^{\mathrm{L}} = 1$ with the color being the benefit value; b) the distribution can be controlled by tuning α and β. Each dot in the curve plane is one optimal solution of playback time $\{s_1, s_2, s_3\}$, with the color being its corresponding α;

order of θ_t, until the time constraint u^{L} is reached, which in fact implements the conventional key-frame extraction (e.g., Eq.1). Only when $0 < \beta < 1$, it forms a concave optimization function, and distributes the playing time well into frames.

We rewrite Eq.4 into an unconstrained form with a Lagrange multiplier γ

$$\hat{\mathbf{V}}^* = \arg\max_{\hat{\mathbf{V}}} \left[\sum_{t=1}^{N} \left[f_t \frac{a_t}{C_t^{\mathrm{c}}} \right]^{\alpha} s_t^{\beta} \right] + \gamma \left(\sum_{t=1}^{N} s_t - u^{\mathrm{L}} \right). \quad (5)$$

By partially differentiating it w.r.t. each s_t and setting it to zero, we derive the optimal solution of s_t under $0 < \beta < 1$ as

$$s_t \propto \left[f_t \frac{a_t}{C_t^{\mathrm{c}}} \right]^{\frac{\alpha}{1-\beta}}. \quad (6)$$

Using the above example, we plot the relationship between u^{L} and its optimal distribution of s_t under different α values in Fig. 1(b). When $\alpha = 1 - \beta$ (e.g., Eq.3), the optimal s_t will be linearly proportional to θ_t, as shown by the three projections on XY, YZ, XZ planes. When $\alpha > 1 - \beta$, the criterion favours assigning higher s_t to frames with a higher θ_t. The higher α, the closer the vertical axis s_3. When $\alpha < 1 - \beta$, the criterion provides more even distribution in all frames. The smaller α, the closer the line $s_1 = s_2 = s_3$ (Fig. 1(b)).

There are two possible choices to specify C_t^{c} in Eq.4 without the need to explicitly know its exact value: 1) Let $C_t^{\mathrm{c}} = a_t$,

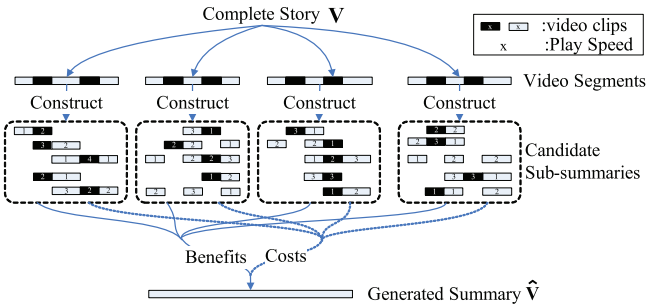


Fig. 2. Conceptual diagram of the overall proposed summarization process envisioned in a divide and conquer paradigm.

which implicitly includes the video stimuli, by assuming that the original video is already optimal in its strength of stimulus; 2) Let $\forall t, C_t^{\mathrm{c}} = C^{\mathrm{c}}$, which forces all frames to have equalized target stimuli, where C_t^{c} is included into the normalization term in Eq.6. The former choice is more suitable for professionally produced videos by experts (e.g., broadcasted videos) but fails in controlling visual stimuli in unedited videos (e.g., surveillance videos with sparse activities). In order to include video stimuli explicitly, we adopt the latter choice.

One major limitation of the criterion in Eq. 4 (also in Eq.1 and Eq.3) is its basic assumption on the linear additivity of information F, which does not always hold (e.g., when handling complicated story-telling with strong internal dependency or considering the emotional involvement of the audiences), and thus constraints its applications. Instead of directly summarizing the video based on Eq.4, we propose a resource allocation framework, which takes Eq.4 as the base criterion but introduces the non-linearity by performing a hierarchical summarization, as discussed in the next section.

IV. Resource Allocation Framework

Our resource-allocation-based framework interprets the summarization problem as finding the optimal allocation of duration resources u^{L} into video segments, according to various user preferences. We design the whole process using the divide and conquer paradigm (Fig.2(a)). The whole video is first cut into short clips by using a shot-boundary detector. These short clips are then organized into video segments. A sub-summary or local story defines one way to select clips within a segment. Several sub-summaries can be generated from a segment: not only the content, but also the narrative style of the summary can be adapted to user requirements. By tuning the benefit and the cost of sub-summaries, we balance in a natural and personal way the semantics (what is included in the summary) and the narrative (how it is presented to the user) of the summary. The final summary is formed by collecting non-overlapping sub-summaries to maximize the overall benefit, under the user-preferences and duration constraint.

This hierarchical framework also helps to overcome the limitation posed by the linear additivity of information/benefit. Each segment is complete in describing an activity/event. The information/benefits of the segments are supposed to be linearly additive. The video clip is our minimum summarization unit, which means that its frames are handled together. Non-linear accumulation of information among clips is processed

with our non-linear local story organization described in Section IV-B. Non-linear accumulation of information within a clip is computed by its benefit as discussed in Section V.

The proposed framework is applicable to any segmented videos with information values for f_t and a_t, independent from the detailed definition and implementations of those two notions. We first discuss the summarization framework by assuming the availability of video segments and information values. Specific methods for video segmentation and information computation will be given along with its application in two use cases in the Section V.

A. Preliminaries

Let the video be cut into N^{C} clips, with the i^{th} clip $\mathcal{C}_i$ being $\mathcal{C}_i = \{t|t = t_i^{\mathrm{S}}, \cdots, t_i^{\mathrm{E}}\}$. t_i^{S} and t_i^{E} are the indices of its starting and ending frames. These video clips are grouped into M segments. A set of candidate sub-summaries is considered for each segment, from which at most one sub-summary is selected into the resulting summary. We denote the k^{th} sub-summary of the m^{th} segment $\mathcal{S}_m$ as $\mathbf{a}_{mk}$, which is a set of playback speeds for all its clips, i.e. $\mathbf{a}_{mk} = \{v_{ki}|i \in \mathcal{S}_m\}$. v_{ki} is the playback speed assigned to the i^{th} clip if the k^{th} sub-summary $\mathbf{a}_{mk}$ is adopted. The summary is then denoted as $\hat{\mathbf{V}} = \cup_{m=1}^{M} \cup_{i\in\mathcal{S}_m} \{(f_t, s_t = 1/v_{ki})|t \in \mathcal{C}_i\}$.

Let $\mathbf{b}_m = \{b_i|i \in \mathcal{S}_m\}$ be the list of base benefits for all clips in $\mathcal{S}_m$. Our major task is to find the set of sub-summaries that maximizes the total pay-off

$$\hat{\mathbf{V}}^* = \arg\max_{\hat{\mathbf{V}}} \mathcal{B}(\{\mathbf{a}_{mk}\}|\{\mathbf{b}_m\}), \tag{7}$$

subject to $\sum_{m=1}^{M} |\mathbf{a}_{mk}| \leq u^{\mathrm{L}}$. We define $|\mathbf{a}_{mk}|$ as the overall length of summary $\mathbf{a}_{mk}$,

$$|\mathbf{a}_{mk}| = \sum_{i\in\mathcal{S}_m} \frac{t_i^{\mathrm{E}} - t_i^{\mathrm{S}}}{v_{ki}}. \tag{8}$$

The overall benefit of the whole summary is defined as accumulated benefits of all selected sub-summaries:

$$\mathcal{B}(\{\mathbf{a}_{mk}\}|\{\mathbf{b}_m\}) = \sum_{m=1}^{M} \mathcal{B}_m(\mathbf{a}_{mk}), \tag{9}$$

with $\mathcal{B}_m(\mathbf{a}_{mk})$ being defined as a function of the user preferences, of the highlighted moments, and of the playback speeds as described in the following.

B. Local Story Organization

One major advantage of the resource allocation framework is that it allows highly personalized story organization, which is achieved via flexible definition of benefits. We define the benefit of a sub-summary as

$$\mathcal{B}_m(\mathbf{a}_{mk}) = \sum_{i\in\mathcal{S}_m} \mathcal{B}_i(v_{ki}) \mathcal{B}_{mi}^{\mathrm{P}}(\mathbf{a}_{mk}), \tag{10}$$

which includes accumulated benefits of selected clips. $\mathcal{B}_i(v_{ki})$ computes the base benefit of clip i at playback speed v_{ki},

$$\mathcal{B}_i(v_{ki}) \quad = \quad b_i(1/v_{ki})^{\beta}, \tag{11}$$

with

$$b_i \quad = \quad \sum_{t=t_i^{\mathrm{S}}}^{t_i^{\mathrm{E}}} (f_t a_t)^{\alpha} \tag{12}$$

being the base benefit of clip i. $\mathcal{B}_{mi}^{\mathrm{P}}(\mathbf{a}_{mk})$ evaluates the extra benefits by satisfying specific preferences:

$$\mathcal{B}_{mi}^{\mathrm{P}}(\mathbf{a}_{mk}) = \mathcal{P}^{\mathrm{O}}(v_{ki}, u^{\mathrm{O}}) \mathcal{P}_{mki}^{\mathrm{C}}(u^{\mathrm{C}}) \mathcal{P}_{mk}^{\mathrm{F}}. \tag{13}$$

$\mathcal{P}^{\mathrm{O}}(v_{ki}, u^{\mathrm{O}})$ is the extra gain obtained by including the user's favourite object u^{O}, specified through an interactive interface,

$$\mathcal{P}^{\mathrm{O}}(v_{ki}, u^{\mathrm{O}}) = \begin{cases} \phi, & v_{ki} < \infty, \exists t \in \mathcal{C}_i, u^{\mathrm{O}} \text{exists in} I_t, \\ 1.0, & \text{otherwise.} \end{cases} \tag{14}$$

$\phi(> 1.0)$ is a parameter to control the strength of emphasizing the favourite object in the summary. We favour a continuous story-telling by defining $\mathcal{P}_{mki}^{\mathrm{C}}(u^{\mathrm{C}})$

$$\mathcal{P}_{mki}^{\mathrm{C}}(u^{\mathrm{C}}) = 1 + u^{\mathrm{C}}(2 - \delta_{\frac{1}{v_{ki}v_{k(i+1)}},0} - \delta_{\frac{1}{v_{ki}v_{k(i-1)}},0}), \tag{15}$$

where $\delta_{a,b}$ is the Kronecker delta function, and u^{C} is fixed to 0.1 in our experiments. Satisfaction of general production principles is also evaluated through $\mathcal{P}_{mk}^{\mathrm{F}}$, which takes 1 for normal case and 0.001 for forbidden cases (or a value that is small enough to suppress this case from being selected), to avoid unpleasant visual/story-telling artifacts (e.g., too-short/incomplete local stories). In summary, the current framework supports user preferences on time duration u^{L}, favourite object u^{O} and story continuity u^{C}.

C. Global Story Organization

The global-duration resource is allocated among the available sub-summaries to maximize the aggregated benefit (Eq.7). Under strict constraints, the problem needs to rely on heuristic methods or dynamic programming to be solved. However, when relaxation of constraints is allowed, Lagrangian optimization and convex-hull approximation can be considered to split the global optimization problem in a set of simple block-based decision problems [39][40]. The convex-hull approximation restricts the eligible summarization options for each sub-summary to the (benefit, cost) points sustaining the upper convex hull of the available (benefit, cost) pairs of the segment. Global optimization is obtained by allocating the available duration among the individual segment convex-hulls [41]. This results in a computationally efficient solution that considers a set of candidate sub-summaries with various descriptive levels for each segment. Fig.3 summarizes the summarization process based on solving a resource allocation problem.

We solve this resource allocation problem by using the Lagrangian relaxation [41]: if λ is a non-negative Lagrangian multiplier and $\{k^*\}$ is the optimal set that maximizes

$$\mathcal{L}(\{k\}) = \sum_{m=1}^{M} \mathcal{B}_m(\mathbf{a}_{mk}) - \lambda \sum_{m=1}^{M} |\mathbf{a}_{mk}| \tag{16}$$

over all possible $\{k\}$, then $\{\mathbf{a}_{mk^*}\}$ maximizes $\sum_{m=1}^{M} \mathcal{B}_m(\mathbf{a}_{mk})$ over all $\{\mathbf{a}_{mk}\}$ such that $\sum_{m=1}^{M} |\mathbf{a}_{mk}| \leqslant \sum_{m=1}^{M} |\mathbf{a}_{mk^*}|$. Hence, if $\{k^*\}$ solves the unconstrained problem in Eq.16, then it also provides the optimal solution

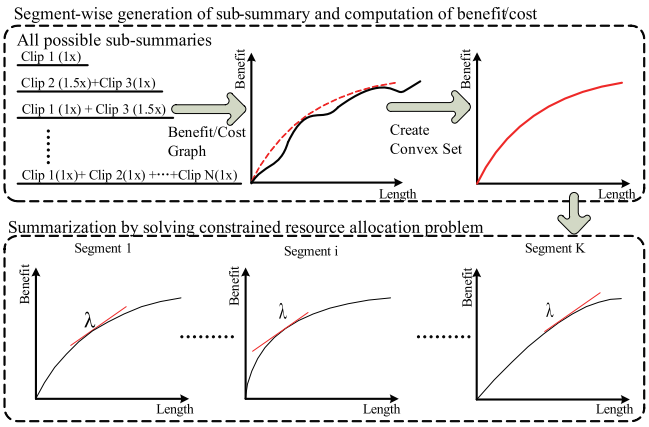


Fig. 3. Lagrangian relaxation and convex-hull approximation are adopted to solve the resource allocation problem, which restrict the eligible summarization options to the convex hulls of benefit-to-cost curves of the segments, where the collection of points from all convex-hulls with a same slope λ produces one optimal solution under the corresponding summary length.

to the constrained problem in Eq.7, with $u^{\mathrm{L}} = \sum_{m=1}^{M} |\mathbf{a}^*_{mk}|$. Since the contributions to the benefit and cost of all segments are independent and additive, we can write

$$\sum_{m=1}^{M} \mathcal{B}_m(\mathbf{a}_{mk}) - \lambda \sum_{m=1}^{M} |\mathbf{a}_{mk}| = \sum_{m=1}^{M} (\mathcal{B}_m(\mathbf{a}_{mk}) - \lambda |\mathbf{a}_{mk}|). \quad (17)$$

From the curves of $\mathcal{B}_m(\mathbf{a}_{mk})$ with respect to their corresponding summary length $|\mathbf{a}_{mk}|$, the collection of points maximizing $\mathcal{B}_m(\mathbf{a}_{mk}) - \lambda|\mathbf{a}_{mk}|$ with a same slope λ produces one unconstrained optimum. Different choices of λ lead to different summary lengths. If we construct a set of convex hulls from the curves of $\mathcal{B}_m(\mathbf{a}_{mk})$ with respect to $|\mathbf{a}_{mk}|$, we can use a greedy algorithm to search for the optimum under a given constraint u^{L}. The approach is depicted in Fig.3 and explained in details in [40]. In short, for each point in each convex hull, we first compute the forward (incremental) differences in both benefits and summary-lengths. We then sort the points of all convex-hulls in decreasing order of λ, i.e. of the increment of benefit per unit of length. Given a length constraint u^{L}, ordered points are accumulated until the summary length gets larger or equal to u^{L}. Selected points on each convex-hull define the sub-summaries for each segment.

Fig.4 shows the clip benefit $\mathcal{B}_i(v)$ w.r.t. $1/v$ under various β and b_i values, so as to analyse the behaviour the clip interest defined in Eq.11 in the above optimization process. Fig.4(a) reveals that the whole curve is convex when $0 < \beta < 1$, which thus enables various options of playback speeds to appear in the benefit/cost convex hulls. In Fig.4(b), we found that the clip with a higher base interest b_i has the same slope value at a slower playback speed. Accordingly, in the above greedy algorithm, slower playback speeds will be first assigned to semantically more important clips in the sense of both high information level and high complexity.

Inclusion of fast-forwarding options significantly increases the number of possible sub-summaries. Compared to [36][37] where naive enumeration of all combinations in a segment is adopted, we consider a sub-optimal way to build the convex hulls. Specifically, we consider the possibility to divide a long segment into shorter sub-segments, and build the convex-hull

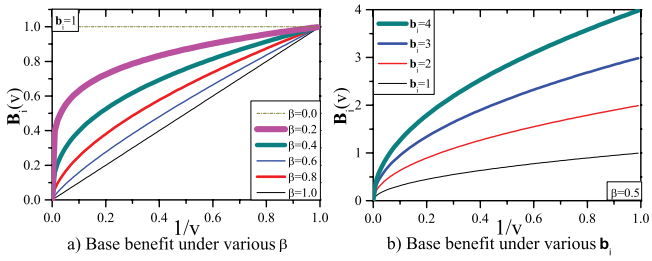


Fig. 4. Clip benefit (Eq.12) also complies with convex-hull approximation and the greedy algorithm adopted for solving the resource allocation problem.

from the convex-hulls of the sub-segments, which provides accurate results when we omit the benefit defined in Eq.13, according to Theorem 4.1. Now we check the terms defined in Eq.13. $\mathcal{P}(v_{ki}, u^{\mathrm{O}})$ is an extra weight computed individually for each clip, which is dividable into sub-segments. $\mathcal{P}^{\mathrm{C}}_{mki}(u^{\mathrm{C}})$ assigns extra weights when consecutive clips are selected, which could be divided into two cases: consecutive clips within each sub-segment are computed first; then connective clips between different sub-segments are considered along with $\mathcal{P}^{\mathrm{F}}_{mk}$ when merging the sub-segments.

Definition Let the benefit-length curve of the m^{th} segment be $\mathcal{B}(x) = \max_{|\mathbf{a}_{mk}|=x} \mathcal{B}_m(\mathbf{a}_{mk})$. Its convex envelop is defined as $\hat{\mathcal{B}}(x)$, which satisfies

- Envelop:

$$\forall x, \hat{\mathcal{B}}(x) \geq \mathcal{B}(x); \quad (18)$$

$$\hat{\mathcal{B}}(x) = \arg\min_{\hat{\mathcal{B}}(x)} \int_x |\hat{\mathcal{B}}(x) - \mathcal{B}(x)| dx; \quad (19)$$

- Convexity: $\forall x_1 < x, x_2 > x$,

$$\hat{\mathcal{B}}(x) \geq \frac{x_2 - x}{x_2 - x_1} \hat{\mathcal{B}}(x_1) + \frac{x - x_1}{x_2 - x_1} \hat{\mathcal{B}}(x_2). \quad (20)$$

A point x^* is called a support point at the convex hull $\hat{\mathcal{B}}(x)$ if it satisfies $\hat{\mathcal{B}}(x^*) = \mathcal{B}(x^*)$.

Theorem 4.1: Assume that we have a dividable benefit function $\mathcal{B}(x)$, i.e. $\mathcal{B}(x) = \mathcal{B}_a(x_a) + \mathcal{B}_b(x_b)$ with $x = x_a + x_b$. If at the support point x^*, we have $\hat{\mathcal{B}}(x^*) = \mathcal{B}(x^*) = \mathcal{B}_a(x^*_a) + \mathcal{B}_b(x^*_b)$, then x^*_a and x^*_b are also support points in both sub-segments.

Proof: Assuming that x^*_a is not a support point in the convex hull of $\mathcal{B}_a(x^*)$, we have $\exists x_{a1} < x^*_a, x_{a2} > x^*_a$,

$$\mathcal{B}_a(x^*_a) < \lambda_a \mathcal{B}_a(x_{a1}) + (1 - \lambda_a)\mathcal{B}_a(x_{a2}), \quad (21)$$

$$\lambda_a = \frac{x_{a2} - x^*_a}{x_{a2} - x_{a1}}. \quad (22)$$

Hence, we have

$$\begin{aligned} \hat{\mathcal{B}}(x^*) &= \mathcal{B}_a(x^*_a) + \mathcal{B}_b(x^*_b) \\ &< \lambda_a \mathcal{B}_a(x_{a1}) + (1 - \lambda_a)\mathcal{B}_a(x_{a2}) + \mathcal{B}_b(x^*_b) \\ &= \lambda_a \mathcal{B}(x_{a1} + x^*_b) + (1 - \lambda_a)\mathcal{B}(x_{a2} + x^*_b) \\ &\leq \lambda_a \hat{\mathcal{B}}(x_{a1} + x^*_b) + (1 - \lambda_a)\hat{\mathcal{B}}(x_{a2} + x^*_b), \end{aligned} \quad (23)$$

which is contradictory to its convexity. Therefore, all support points in the convex hull $\hat{\mathcal{B}}(x)$ must be constructed from support points in the convex hulls $\hat{\mathcal{B}}_a(x_a)$ and $\hat{\mathcal{B}}_b(x_b)$. ■

V. Use Cases

We focus on two use cases: the summarization of unedited videos captured by fixed cameras (surveillance); and the summarization of produced contents with moving cameras and various shot types (broadcasted sport). Unlike previous methods that considered low-level features only, e.g., motion vectors [9] or frame differences [10], we consider video tracking and hot-spot detection on surveillance videos, and the combination of player tracking with detection of camera motions and various production actions for processing broadcasted soccer videos.

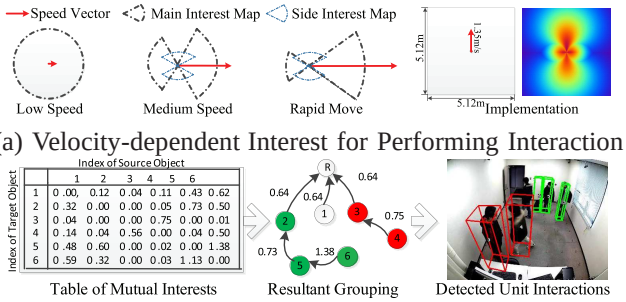
A. Summarization of Surveillance Videos

As video surveillance aims to monitor the activities of objects in the scene, the larger the number of moving objects, the more relevant the scene is expected to be; with equal number of objects, the closer the objects, the slower the playback speed should be. We are thus motivated to link group interactions, defined as stable and continuous spatial proximity between objects, to the adaptive fast-forwarding. Assuming that all objects intend to keep their individual moving status as long as possible [42], group interactions also provide cues to locate spatial-temporal hot-spot events, which facilitates the clip division and video segmentation as well as assignment of clip benefits. We detect group interactions from trajectories extracted by video tracking.

Let us denote the object on the l^{th} trajectory at the t^{th} frame with $\mathbf{o}_{lt} = \{a_{lt}, \mathbf{x}_{lt}\}$. a_{lt} is for the availability of a trajectory, which takes 1 when it appears in the present frame and takes 0 otherwise. $\mathbf{x}_{lt}$ is its position. At the t^{th} frame, we group all moving objects as $\mathcal{G}_t = \{\mathbf{o}_{lt} | a_{lt} = 1\}$. We assume that the movement of each object is driven by the intention to interact with other objects, and define his interest in interacting with an object at position $\mathbf{x}$ as a velocity-dependent function $\mathcal{I}_{lt}(\mathbf{x})$ shown in Fig.5(a). The group interaction is then defined as the behaviour of multiple objects motivated by unidirectional/mutual interests, and is modelled by a directed graph, with the edges being the mutual interests, as shown in Fig.5(a). For objects having no high interests on other objects, we simply let it focus on a virtual object $\mathbf{o}^{\text{V}}$ with fixed interest $\mathcal{I}_{\text{T}}$. Limiting each object to mainly focus on only one target object, we solve the object grouping in each frame by finding the spanning tree of this graph with the maximum interests. Inspired by online object tracking, we obtain group interactions in three steps: Grouping objects into unit interactions at each frame; temporal association of unit interactions; and refinement of detected interactions by post-smoothing [43].

Let g_{lt} be the index of the group interaction that $\mathbf{o}_{lt}$ belongs to and $g_{lt} = 0$ if $\mathbf{o}_{lt}$ is not joining any interaction. At the t^{th} frame, we form a L-dimensional vector for all the L trajectories $\mathbf{I}_t = [\mathbf{I}_{tl} | l = 1, \cdots, L]$, where $\mathbf{I}_{tl}$ is the overall interest it receives from all interacting neighbours

$$\mathbf{I}_{tl} = \begin{cases} a_{lt}\left[1 + \sum\limits_{\mathbf{o}_{mt} \in \mathcal{G}_t \setminus \mathbf{o}_{lt}}^{g_{mt}=g_{lt}} a_{mt}\mathcal{I}_{mt}(\mathbf{x}_{lt})\right], & g_{lt} \neq 0, \\ a_{lt}, & g_{lt} = 0. \end{cases} \quad (24)$$



(a) Velocity-dependent Interest for Performing Interaction

(b) Graph-theory based Extraction of Unit Interaction

Fig. 5. Continuous spatial proximity among moving objects is extracted as group interactions. a) When moving faster, an object gets preferred directions of interaction; b) We model the mutual interest among multiple objects with a graph, and extract the units of interacting objects by finding the maximum weight spanning tree.

We cut the video into short clips at the boundaries of group interactions, and then group clips containing the same interactions as a segment. The two kinds of information in the t^{th} frame are defined as

$$f_t = |\mathbf{I}_t|, \quad (25)$$

$$a_t = |\mathbf{I}_t - \mathbf{I}_{t-1}|. \quad (26)$$

B. Summarization of Broadcasted Soccer Videos

We divide the soccer video into clips, according to the detected production actions, such as position of replays, shot-boundaries and view types. Instead of using (complex) semantic scene analysis tools, we segment the video based on the monitoring of production actions by analysing the view-structure [37]: We detect replays from producer-specific logos [44], extract shot-boundaries with a detector proposed in [45] to better deal with smooth transitions, and recognize the view-type by using the method in [26]. As in [36], we automatically locate hot-spots by analysing audio signals [46], whose (change of) intensity is correlated to the semantic importance of each video segment. We consider the average information associated to still contents $\overline{f_t}$ and that associated to scene changing $\overline{a_t}$ evaluated on the clip level. Accordingly, we compute the approximated form of clip benefit in Eq.12,

$$b_i = |t_i^{\text{E}} - t_i^{\text{S}}| \left(\overline{f_t}\,\overline{a_t}\right)^{\alpha}. \quad (27)$$

Beyond a chronological and complete (using far views) presentation of the game, the professionals also attempt to involve the audience emotionally by highlighting the dominant player with close-up views and emphasizing the most exciting moment with replays [47]. The benefit of each frame t within each segment is thus evaluated from its relevance to the game f_t^{G} and its level of emotional involvement f_t^{E}. The frame information f_t is computed as

$$f_t = 0.25 f_t^{\text{E}} + 0.75 f_t^{\text{G}}. \quad (28)$$

We use the above fixed weight to favour game related contents in the summary. In practice, it is very complicated to define the f_t^{E} and f_t^{G} metrics. This could for example be done by identifying the dominant player from a set of consecutive close-up views or by confirming the replay to its corresponding far-view clips taken at different camera positions. Instead, we consider an heuristic approach that roughly distributes the importance of detected hot-spots into the clips in a segment

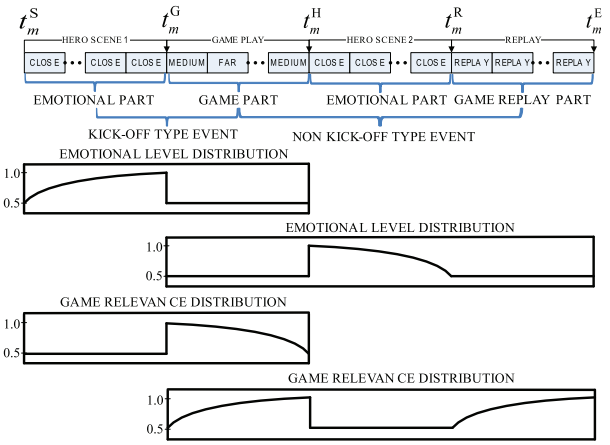


Fig. 6. The base benefit of a clip is evaluated from the game relevance and emotional level, defined as functions of clip view-types. The decaying process is modelled by hyperbolic tangent function. t_m^{G}, t_m^{H}, t_m^{R} are starting times of game play, hero scene, and replay in the m^{th} segment, respectively.

based on the general production rules: The dominant player is usually the last to be presented before an important action, but the first to be shown after an action; The close-up views and replays are usually inserted right after an important action, which suggests that the closer a far view is to the close-up view or the replay clip, the more relevant it is [47]. Hence, we define f_t^{G} and f_t^{E} by propagating the significance of the detected hot-spot event according to the view type structure of the segment, as depicted in Fig.6. The decaying process was modelled by using the hyperbolic tangent function, because it is bounded and is integrable, thus simplifying the computation of $\overline{f_t}$. Since our allocation of resources directly depends on the proposed model, our experimental results tend to confirm the relevance of the adopted model indirectly via the subjective assessment of users satisfaction about the generated summaries. Note that if a more accurate model was developed regarding the emotional and game interest of a video, e.g., based on the affective computing literature [48][49], it would be straightforward to integrate it within the framework, as long as the model assigns benefits in a way that is additive over video segments (i.e. the benefit associated to a segment is independent from other segments).

Information associated to scene changing a_t is defined on the fluctuation of the camera view or the diversified movement of multiple players. Given a clip, the fluctuation of its camera view $\overline{\tau^{\mathrm{M}}}$ is evaluated by the average standard deviation of the motion vectors in the clip, while the complexity of diversified player movements $\overline{\tau^{\mathrm{P}}}$ is defined as the average standard deviation of players' moving speeds in the clip. As shown in Fig.7, the average information $\overline{a_t}$ is then defined as a weighted sum of the above two terms,

$$\overline{a_t} \propto \begin{cases} \overline{\tau^{\mathrm{M}}} + \overline{\tau^{\mathrm{P}}}, & \text{far view} \\ \overline{\tau^{\mathrm{M}}}, & \text{otherwise} \end{cases} \tag{29}$$

which is normalized to $[0 \quad 1]$ for far-view and non-far-view clips independently. Using the standard deviation avoids the need of accurate compensation of player speed with respect to camera motions.

We only allow normal speed for a replay clip in local story

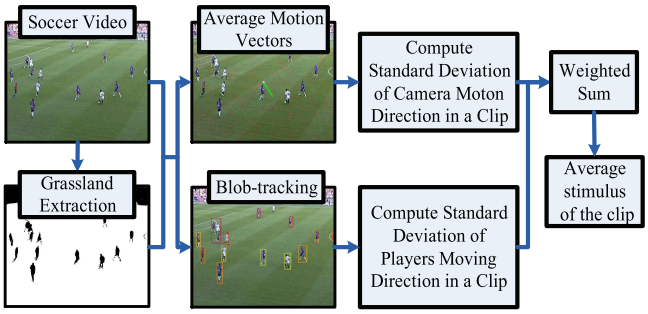


Fig. 7. We evaluate the average stimulus in a far-view clip by estimating information associated to scene activity from camera motion and player motion, which are computed on average motion vector in the grassland region and tracked player positions.

organization. If time resources to render a replay are available, we present the action in the clearest way.

VI. Experimental Results

A. Experimental Setup

We use a broadcasted soccer video and two surveillance videos to validate the performance of our framework. The soccer video is 3 hours long with a list of 50 automatically extracted audio hot-spots. The two surveillance videos include a 7-minute indoor surveillance video from the JAIST dataset [50] and a 14-minute outdoor surveillance video from the Behave dataset [51], both with various group activities between multiple persons. Seven different speed options, i.e. 1x, 2x, 4x, 6x, 8x, 10x, and $+\infty$ (for content truncation[1]), are enabled in the current implementation, so as to provide comparative flexibility in fast-forwarding control to those methods with continuous playback speeds. Here, ax stands for the a times of the normal playback speed. In the multi-view JAIST dataset, we performed conventional tracking after detection methods and achieved accurate ($\sim 95\%$) tracking results [52][53]. Detailed quantitative results and demo videos can be found in [54]. In the single view Behave dataset, we use the trajectories provided by the dataset, where many conventional tracking methods are also available [55].

The proposed framework aims at focusing on summarization with adaptive fast-forwarding and semantically relevant and personalized story telling. Its performance is explored through a comparative analysis with state of the art methods. Especially, we compared the behaviour of our proposed method to three methods, i.e. *Peker et al.* [9], *Höferlin et al.* [10] and *Naive fast-forwarding*.

Peker et al. [9] achieve adaptive fast-forwarding via constant activity sub-sampling

$$s_t^* = \frac{r_t}{r^{\mathrm{target}}} s_t, \tag{30}$$

where the complexity of activity r_t is measured by average motion vector magnitude. We estimated the motion vector by the Horn-Schunck method as originally applied by Peker et al., and used the implementation in OpenCV.

[1]When content truncation is not desired in some surveillance systems, we could also replace $+\infty$ with the maximum fast-forwarding playback speed allowed in the deployed system, e.g., 64x.

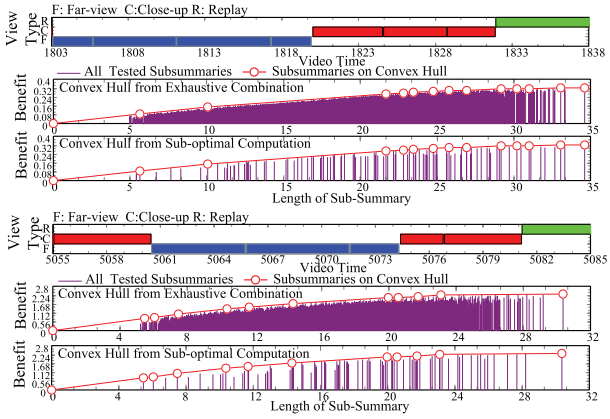
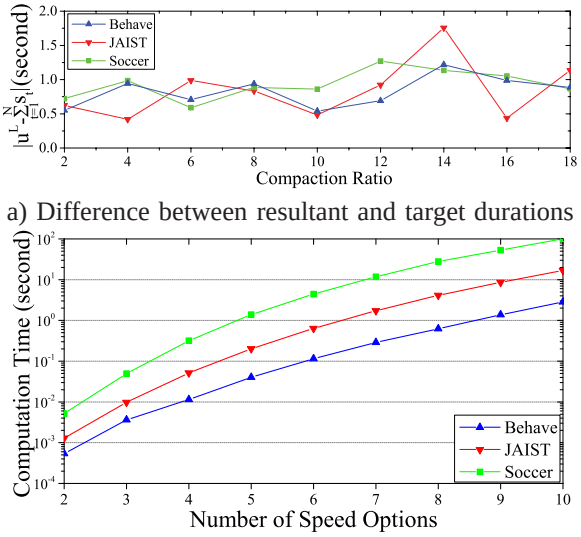


Fig. 8. Compared to exhaustive combination, our approximated computation provides the same convex hulls on two tested segments, with significantly less tested combinations. The top of each sub-figure gives the view structure of the segment along with the clip boundaries. Each vertical bar in the middle and bottom of the sub-figure represents one considered combination. The resultant convex hull is marked in the red curve, and the support sub-summaries in red circles. Five speed options are considered, namely, 1x, 2x, 4x, 8x, and $+\infty$.



a) Difference between resultant and target durations

b) Computational time w.r.t. numbers of speed options

Fig. 9. Behaviours of our proposed method in terms of optimality and computational cost.

Höferlin et al. [10] determine the activity level by computing the alpha-divergence between the luminance difference of two consecutive frames and the estimated noise model. A bigger divergence value stands for a larger distance between the current frame difference and those caused by the background noise. The adjusted sampling interval s_t^* is then set to be linearly proportional to the activity level. We learnt the noise model from several training clips of background scenes without moving foreground objects and camera motions. Alpha was set to 1, which results in the Kullback-Leibler divergence and was most discussed in [10].

Naive fast-forwarding simply assigns a uniform playback speeds to all frames.

We only provide representative results directly related to the summarization performance here. The corresponding videos and additional experimental results are available in the supplemental materials associated to this paper [56].

B. Behaviour of the proposed method

In Fig.8, we compared convex-hulls of sub-summaries from exhaustive combination and our approximated computation on two long segments with multiple clips. When computing the convex-hull for a segment with C clips (where each clip could take S different speeds), we have S^C different combinations in exhaustive enumeration. If we divide the long segment into short sub-segments of C_2 clips, we only need $(C/C_2)(S^{C_2}+K^2)$ times of enumerations. K is the average number of support points in a convex-hull, which empirically is around 20 when $C_2 \leq 7$. The approximated computation provided the same convex hull as the exhaustive combination, with significantly less tested combinations, which is used in the following experiments.

Lagrangian relaxation provides optimal solutions when the generated summary duration is equal to the user-imposed duration constraint u^{L} [40]. We evaluate the potential sub-optimality induced by Lagrangian relaxation by investigating the difference between the length of the resultant summary to its target duration, i.e. $|u^{\mathrm{L}} - \sum_{t=1}^{N} s_t|$, in Fig.9(a) (averaged over different α values). Since the durations of the summaries generated based on convex-hull operating points are close to the constraint (with averaged difference around 1s), the sub-optimality is negligible.

Since meta-data collection can be performed off-line as preprocessing, we mainly discuss the computation cost in producing the summary, which is more relevant in online summarization service of pre-recorded videos. Fig.9(b) shows the computational time for summarizing the three videos by a single threaded implementation running on a Core i7 CPU (2.3Ghz), under speed options varying from 2 to 10. For a short segment of C_2 clips, increasing the number of speed options by one slows the enumeration process by $[1+1/S]^{C_2}$, which gradually saturates to 1 when S increases. When $C_2 = 5$ and $S \geq 6$ (i.e. $[1+1/S]^{C_2} < 2.17$), the overall computational time almost doubles when one playback option is added (Fig.9(a)), i.e. the approximated computation successfully linearized the computation between short segments. The computation of normalized inverse linear proportion in [9] and [10] costs about 2ms (JAIST: 1.62 ± 0.20ms, Behave: 2.31 ± 0.30ms and Soccer: 2.46 ± 0.03ms. Averaged after 20 trials). Although slower, the proposed method can still be regarded as real-time responsive, if the viewers can get the generated summary in $1 \sim 2$ seconds after inputting their preferences, according to the limits of response times found in usability engineering [57]. Note that the computation can be further accelerated by parallel optimization of the local story organization in different segments, which is a straightforward extension in our divide and conquer framework.

C. Objective Evaluation

The summaries for objective evaluation are generated from the whole videos of both the JAIST and Behave Datasets and the period of $1020s$-$2030s$ in the soccer video, by varying the compaction ratio (defined as N/u^{L}) from 2 to 20. We denote the set of ground truth events as $\mathbf{E}^{\mathrm{GT}} = \{\mathbf{e}_q^{\mathrm{GT}} | q = 1, \cdots, N^{\mathrm{GT}}\}$. Each event has three elements,

$\mathbf{e}_q^{\mathrm{GT}} = (\tau_q^{\mathrm{GT}}, \mathcal{C}_q^{\mathrm{GT}}, \mathcal{G}_q^{\mathrm{GT}})$, corresponding to its type, temporal period and related member objects. Let θ_q be the importance value of the q^{th} event. The ground-truth includes 27 events for the soccer video, 51 events for the JAIST video and 52 events for the Behave video, which are classified into four tiers according to their relative importance (Table II). We compare the above methods with multiple objective criteria for investigating the following behaviours:

1) *Adaptive fast-forwarding for semantic event browsing*. Given the summary $\hat{\mathbf{V}} = \{(I_t, s_t)|t = 1, \cdots, N, s_t \in [0 \quad 1]\}$, we define the first criterion L_1 as the normalized information density of its frames

$$L_1 = \frac{\sum_{t=1}^{N} s_t \sum_{q=1}^{N^{\mathrm{GT}}} \psi_q(t)\theta_q}{\sum_{t=1}^{N} \sum_{q=1}^{N^{\mathrm{GT}}} \psi_q(t)\theta_q} / \frac{\sum_{t=1}^{N} s_t}{N}, \quad (31)$$

which is plotted in Fig.10(a). $\psi_q(t)$ determines whether the q^{th} event occurs at the t^{th} frame.

$$\psi_q(t) = \begin{cases} 1, & t \in \mathcal{C}_q^{\mathrm{GT}} \\ 0, & \text{otherwise.} \end{cases} \quad (32)$$

Both [9] and [10] obtain low L_1 values, which suggests that they failed to correctly measure the intensity of scene activities or the importance of the events. In the soccer video, grasslands in the far view lead to motion vectors of lower magnitude and less noticeable frame differences. Since the events are annotated on far-view clips, the fact that [9] and [10] have even lower L_1 values than the naive fast-forwarding suggests that more time resources are allocated to close-up views[2], although close-up views are reported to tolerate higher playback speeds than far-views in the subjective tests presented in Fig.11. For surveillance videos without camera motions, both the optical flow and the alpha divergence become less sensitive in reflecting the activities in the scene. In contrast, our method achieves higher L_1 values than other methods, which shows that the proposed method is more semantically relevant to the annotated events, by assigning slower playback speeds to clips with both higher event importance and scene activities.

2) *Adaptive fast-forwarding for visually comfortable summarization*. A comfort summary need to be played back slowly enough (supported by the subjective tests presented in Section VI-D), and the speeds should vary gradually so as to avoid annoying flickering. The comfort is evaluated by both the average playback speed L_2 and the fluctuation level of playback speeds between consecutive frames L_3. We consider the non-truncated content, i.e. sub-sequence $\hat{\mathbf{V}}^* = \{(I_t, s_t)|t = 1, \cdots, N^*; s_t > 0\} \subset \hat{\mathbf{V}}$, and define L_2,L_3 as

$$L_2 = E[1/s_i] = \frac{N^*}{\sum_{t=1}^{N^*} s_i}, \quad (33)$$

$$L_3 = \sigma[\Delta(1/s_t)] = \sqrt{\frac{\sum_{t=2}^{N^*} s_t (\frac{1}{s_t} - \frac{1}{s_{t-1}})^2}{\sum_{t=2}^{N^*} s_t}}, \quad (34)$$

which are shown in Fig.10(b)(c). When the length of target summary changes, playback speeds of different clips in [9] and [10] maintain the same ratio. Accordingly, under a high compaction ratio, all clips will be rendered with intolerable speeds. Furthermore, the high fluctuation level L_3 in [9] and [10] stands for frequent and severe playback speed changes in the summary. In contrast, our proposed method is able to maintain a lower playback speed L_2 by truncating the less important contents and has much lower fluctuation level L_3 because of clip-based summarization.

[2]This is confirmed by the plotted distribution of the playback time and the highlight curve in the supplemental material.

TABLE II
MANUALLY ANNOTATED EVENTS AND RELATIVE IMPORTANCE

Tier	Soccer	JAIST	Behave	θ_q
1	Goal	Fight,StealBag	Fight	4
2	Foul Shoot	FallDown ExchangeBag	Split Approach	3
3	PlaceKick, Corner Clearance	DragBag,Pickup StopAndChat	RunTogether InGroup	2
4	Kickoff BallBacktoCourt BallOutofCourt	Following	WalkTogether	1

3) *Adaptive fast-forwarding for narrative story organization*. Compared to the linear playback speed control in [9] and [10], our framework allows flexible personalization of story organization by tuning the time duration u^{L} and the controlling parameters (α, β) (Eq.6). We can suppress redundant contents in the replays for higher compaction, consider story continuity, and remove very short clips to avoid flickering. Our framework can further satisfy the user preferences on favourite objects/events. We define L_4 as the normalized density of information related to a specified object in the summary

$$L_4 = \frac{\sum_{t=1}^{N} s_t \sum_{q=1}^{N^{\mathrm{GT}}} \chi_q(t, u^{\mathrm{O}})\theta_q}{\sum_{t=1}^{N} \sum_{q=1}^{N^{\mathrm{GT}}} \chi_q(t, u^{\mathrm{O}})\theta_q} / \frac{\sum_{t=1}^{N} s_t}{N}, \quad (35)$$

and plot L_4 of the summaries of the JAIST video under various ϕ values (Eq.14) and compaction ratio 8 in Fig.10(d). $\chi_q(t, u^{\mathrm{O}})$ determines whether object u^{O} is involved in the q^{th} event at the t^{th} frame.

$$\chi_q(t, u^{\mathrm{O}}) = \begin{cases} 1, & t \in \mathcal{C}_q^{\mathrm{GT}}, u^{\mathrm{O}} \in \mathcal{G}_q^{\mathrm{GT}} \\ 0, & \text{otherwise.} \end{cases} \quad (36)$$

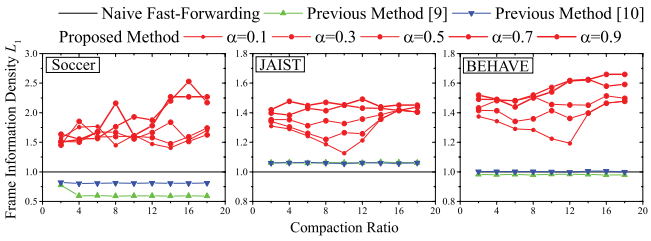
When an object is specified, higher weights will be assigned to its related clips, by results in a larger L_4 value.

D. Subjective Evaluation

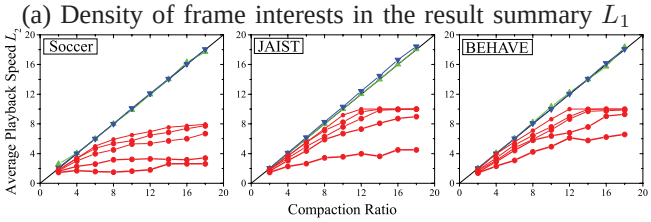
The purpose of our subjective evaluation test is not limited to comparing the performances of the methods, but also to explore possible future improvements through detailed inspection of unnatural story/visual artifacts in the summarization results. Accordingly, we have designed and performed three subjective tests to collect the related opinions of the audiences.

1) The first subjective evaluation evaluates the suitable *playback speeds* (Fig.11). 25 participants (including 11 females and 14 male, age from 20-40) were asked to specify their highest tolerable playback speed, comfortable playback speeds and the most comfortable playback speed when presented five groups of video samples from both broadcasted soccer videos and surveillance videos with various playback speeds.

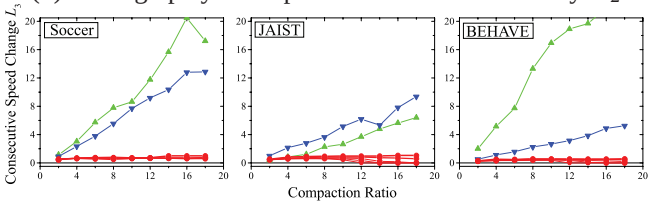
For the soccer video, the highest tolerable speed for far views is lower than that of the close-up views. We consider this as a result that understanding far-view need attentional



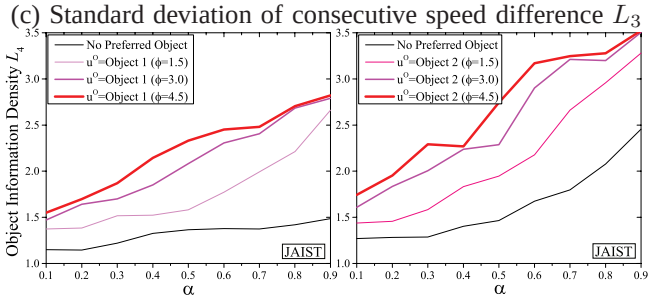
(a) Density of frame interests in the result summary L_1



(b) Average playback speed in the result summary L_2

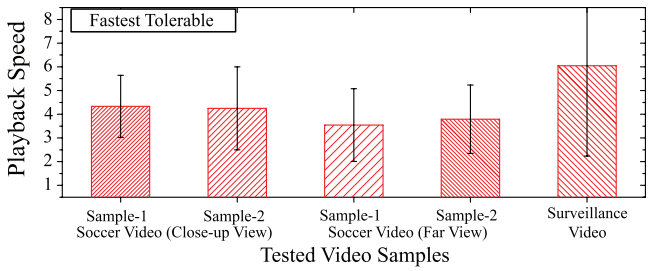


(c) Standard deviation of consecutive speed difference L_3

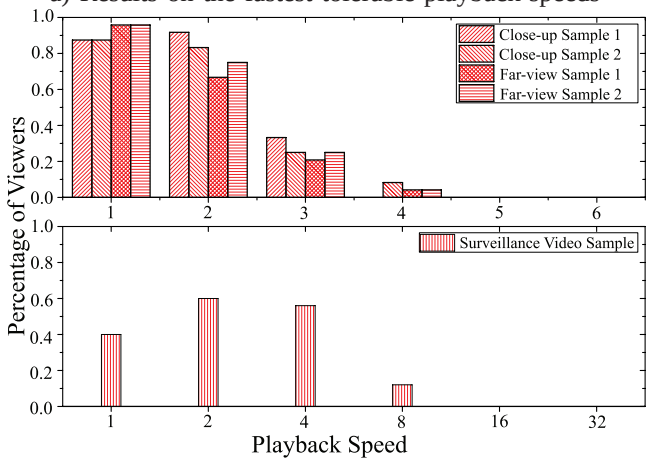


(d) Density of object interests in the result summary L_4

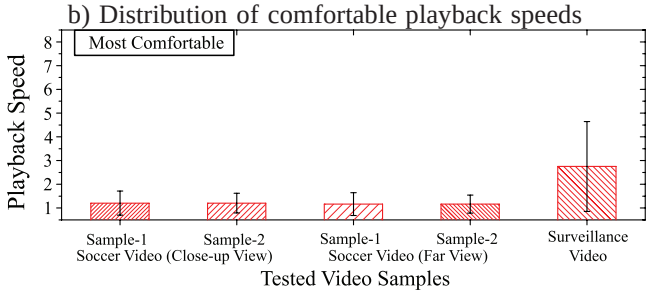
Fig. 10. We plot the results of multiple criteria for objective evaluation of the behaviour of the proposed system ($\beta = 0.5$).



a) Results on the fastest tolerable playback speeds



b) Distribution of comfortable playback speeds



c) Results on the most comfortable playback speeds

Fig. 11. Results of subjective evaluation from 25 participants on their feedback under various fast-forwarding speeds when browsing five video samples from both broadcasted soccer video and surveillance video. The six playback speed options for soccer videos are 1x, 2x, 3x, 4x, 5x and 6x, while that for surveillance videos are 1x, 2x, 4x, 8x, 16x and 32x.

perception to follow the players. For surveillance video, it could tolerate even higher speed, mainly because the fixed camera view makes the selective perception much easier. Most participants cannot tolerate a speed over 4x (i.e. 6FPS in a 25FPS video), which coincides with the observations in previous researches that perception on higher-order motion, word recognition, acceleration/direction change will require a playback speed around or even below 8 FPS [11]. In both cases, audiences still feel comfortable in faster playback speeds, which is the base of adaptive fast-forwarding. As for the most comfortable speed, most audiences prefer the original speed selected and produced by experts in the soccer video. For surveillance video, audiences prefer a faster playback speed (2x or 4x), due to low stimuli in the original video.

2) The second subjective test collects the *global impression* of audiences in comparatively evaluating the generated summaries. We asked 23 participants (including 10 females and 13 males, age from 20-40) to give their opinions on the preferred result when presented a group of three summaries generated by the methods under analysis (in random order) for *completeness*, *comfort*, and *effectiveness* of time allocation. We plot the results of evaluating six-groups of summaries from three source videos under two different compaction ratios (i.e. 8 and 4) along with the questions in Fig.12. Besides the overall conclusion that our method performs the best especially under the high compaction ratio (8), we observed that:

a) Our method outperforms the other two methods in generating complete summaries for highly compact summarization (8), which supports our idea of introducing content truncation to save time resources for presenting key events in a clearer way. With the lowering of the compaction ratio, the average playback speed becomes tolerable or even comfortable, where the viewers could realize the existence of truncated contents and assign a lower completeness value to our method, which is considered to be the reason why [9] outperforms our method in summarizing the Behave dataset under compaction ratio 4.

b) Our method produces more comfortable summaries from the broadcasted soccer video, where both 8 and 4 are too high for an adaptive fast-forwarding method to produce a comfortable video without truncating some contents. In order to slow down a key event, we have to raise the playback speed of other contents to a much higher level in exchange

for the equivalent time resource, which results in flickering and lowers the visual comfort of the summary. Our method also outperforms the other two methods in summarizing the JAIST video, where the close and dense group activities in the scene make the evaluation easier. The difference is less obvious in the Behave dataset due to two major reasons: i) The activities in the video are sparse and simpler; ii) We did not tell the viewers our definition of key-events in order to avoid a biased evaluation towards group-interaction events. The Behave dataset recorded some movements of cars, bicycles and irrelevant pedestrians without providing the corresponding trajectories, which might have distracted the viewers' attentions.

c) Our method is evaluated to be the most effective in allocating playback speeds for presenting the actions of interest, especially under a high compaction ratio.

d) Although [10] was a recent method proposed for summarizing surveillance videos, it fails to outperform [9], especially in summarizing the Behave dataset, mainly due to the difficulty in learning the noise model. Although we have prepared neat training video clips for noise estimation which include no foreground activities and are close to the testing video in terms of lighting conditions, both the noise in the JAIST dataset captured indoor with full HD cameras or the insignificant foreground activities in the Behave dataset captured from a far viewpoint through the window could cause a large bias to the alpha-divergence.

3) The third subjective evaluation is based on a detailed inspection of the generated summaries. Each viewer is asked to point and click via an interface to any kind of visual or story-telling artifacts. The timestamp of clicking is automatically recorded by the tool. We do not ask viewers to input detailed comments after each clicking, because interruption during video playing might distract viewers from focusing on the story evolving in the summary, which should especially be avoided for better evaluating the optimal fast-forwarding speed. As a consequence, we have to find out the reason behind each clicking by analysing the aggregation of clickings, a posteriori. We estimate the density of clickings at each video time by using the Parzen-window function to compensate the delay between the occurrence of story artifacts and the corresponding clicking, where a rectangular window of width 2 seconds is applied to the left side of each clicking. Note that the proposed resource allocation framework does not depend on user clicking for adaptive fast-forwarding (and video skipping). We collect data from 16 participants (including 5 females and 11 males, age from 20-40) and plot them in Fig.13. In each sub-figure, we present the view-structure and the allocated playback speed of the generated summary on the top with the vertical bars for pointing out the positions of content truncations. In the bottom, we give the number of viewers who sensed an artifact at each moment. As an overall evaluation, there is only one artifact that received the recognition of more than half of all viewers in all the three tested summaries, which partially proves that the proposed method could provide visually comfortable summaries to satisfy most of the audiences.

We divide artifacts labelled by more than 1/3 of reviewers into three groups: i) Those correspond to a moment with both a high playback speed and content truncation, including

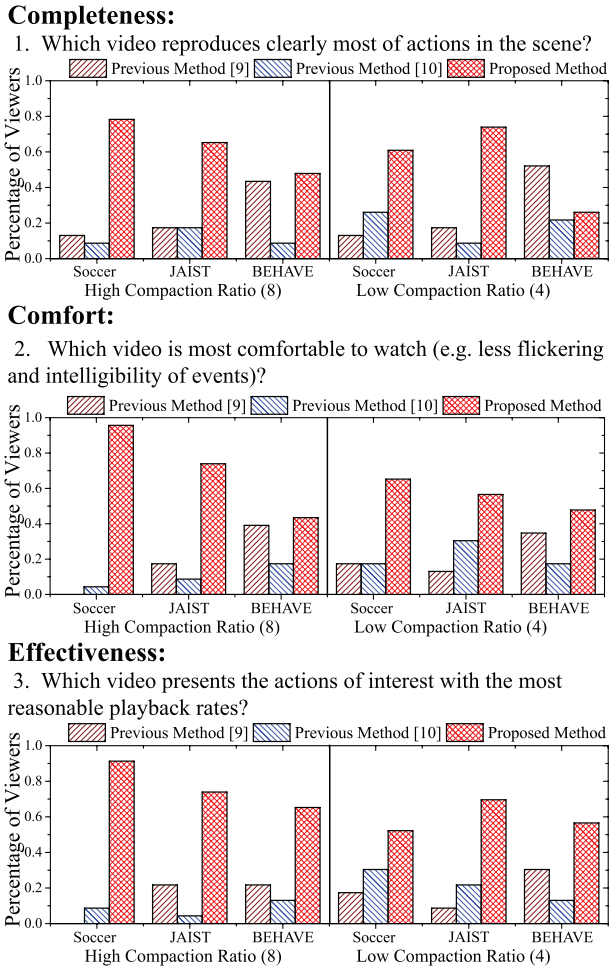
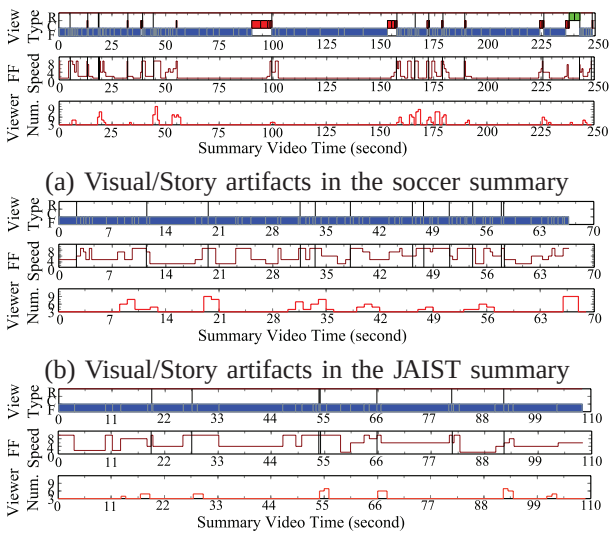


Fig. 12. Results of the second subjective evaluation test from 23 viewers, by collecting their global impression on the summaries, in the sense of completeness, comfort and effectiveness of time allocation.



(a) Visual/Story artifacts in the soccer summary

(b) Visual/Story artifacts in the JAIST summary

(c) Visual/Story artifacts in the Behave summary

Fig. 13. Labelled visual/story-telling artifacts in the third subjective evaluation test by 16 viewers. We present the view structure and the playback speed in the top part of each sub-figure, where the vertical bars present the position of content truncations. In the bottom, we show the aggregated times of artifacts labelled.

21s,48s,170s in soccer video, 20s, 30s and 34s in the JAIST video, and 54s, 66s and 92s in the Behave video; ii) Those correspond to a moment with only a high playback speed, including 55s, 177s in the soccer video, 9s, 40s, 55s and 67s in the JAIST video; iii) Those correspond to a moment with only content truncation, including 225s in the soccer video and 92s in the Behave video. We have the following observations:

a) The viewers are more sensitive to high playback speed than to content truncation, given the fact that most of the above artifacts are related to high playback speeds. We are not surprised with the result, because the playback speed in those artifacts is higher than the comfortable speed revealed in our preliminary subjective evaluation in Fig.11. However, this suggests that content truncation could provide more comfortable summaries than fast-forwarding with a over-fast playback speed, which reinforce our conviction that hybrid summarization with both content truncation and fast-forwarding is the path to follow in the future. In a real application, we could remove these artifacts by limiting the playback speed options within the tolerable range.

b) We notice that clips of high playback speeds usually gather around content truncations. Important clips usually locate in the middle of a segment with neighbouring clips, which is intentionally designed to assure the continuity and completeness of story-telling. We intend to suppress those artifacts in Group 1 by truncating the clips with over-fast playback speeds, and inserting a fixed length transition clip to help the audiences to reorient themselves after truncation.

VII. Conclusions

We proposed a framework for producing personalized summaries that enables both content truncation and adaptive fast-forwarding. We regard adaptive fast-forwarding as a process to tune the stimuli during the information transferring in video browsing, which is important to generate visually comfortable summaries. The limitation of visual perception on the maximum tolerable playback speeds motivated us to consider the hybrid of content truncation and adaptive fast-forwarding to reach a better balance between temporal compression ratio and comfort. Instead of a rigid determination of the fast-forwarding speed, we efficiently select the optimal combination from candidate summaries, which is solved efficiently as a resource-allocation problem. Subjective experiments demonstrate the proposed system by evaluating summaries from both surveillance videos and broadcasted soccer videos.

The proposed framework has the following advantages: 1) higher temporal compression is achievable by increasing the playback speeds to host more content while preserving story-telling continuity; 2) both semantic relevance and visual comfort of the summary are considered by including information associated to still content and scene activity; 3) playback speeds are maintained under a tolerable level by naturally including content truncation in the adaptive fast-forwarding framework; 4) flexible personalization of story-telling is allowed by enabling non-linear story organization in a hierarchical summarization process.

The subjective tests also highlight the direction of further improvements. The audiences could feel comfortable under a faster playback speed, which supports our fast-forwarding based summarization. A too-fast playback speed is found to be even more annoying than content truncation, which drives us to further extend our hybrid method of content truncation and adaptive fast-forwarding. Both information associated to the still contents and scene activity are important in producing a semantically relevant and visually comfort summary. We will thus consider both types of information in our future work.

References

[1] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, no. 1, 2007.

[2] A. Ferman and A. Tekalp, "Two-stage hierarchical video summary extraction to match low-level user browsing preferences," *IEEE Trans. Multimedia*, vol. 5, no. 2, pp. 244–256, 2003.

[3] G. C. de Silva, T. Yamasaki, and K. Aizawa, "Evaluation of video summarization for a large number of cameras in ubiquitous home," in *ACM MM'05*, 2005, pp. 820–828.

[4] Y. Takahashi, N. Nitta, and N. Babaguchi, "Video summarization for large sports video archives," in *ICME'05*, 2005, pp. 1170–1173.

[5] G. Zhu, Q. Huang, C. Xu, Y. Rui, S. Jiang, W. Gao, and H. Yao, "Trajectory based event tactics analysis in broadcast sports video," in *ACM MM'07*, 2007, pp. 58–67.

[6] J. Wang, C. Xu, E. Chng, L. Duan, K. Wan, and Q. Tian, "Automatic generation of personalized music sports video," in *ACM MM'05*, 2005, pp. 735–744.

[7] B. Tseng and J. Smith, "Hierarchical video summarization based on context clustering," in *Internet Multimedia Management Systems IV. (Edited by Smith, J.R.; Panchanathan, S.; Zhang, T.) Proceedings of the SPIE*, vol. 5242, 2003, pp. 14–25.

[8] Z. Li, G. M. Schuster, and A. K. Katsaggelos, "Minmax optimal video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, pp. 1245–1256, 2005.

[9] K. A. Peker, A. Divakaran, and H. Sun, "Constant pace skimming and temporal sub-sampling of video using motion activity," in *ICIP'01*, vol. 3, 2001, pp. 414–417.

[10] B. Höferlin, M. Höferlin, D. Weiskopf, and G. Heidemann, "Information-based adaptive fast-forward for visual surveillance," *Multimedia Tools Appl.*, vol. 55, pp. 127–150, 2011.

[11] A. O. Holcombe, "Seeing slow and seeing fast: two limits on perception," *Trends in Cognitive Sciences*, vol. 13, no. 5, pp. 216 – 221, 2009.

[12] J. Palmer, "Attentional limits on the perception and memory of visual information," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 16, pp. 332–350, 1990.

[13] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *ICIP'98*, vol. 1, 1998, pp. 866–870.

[14] M. Cooper and J. Foote, "Discriminative techniques for keyframe selection," in *ICME'05*, 2005, p. 4.

[15] J. Lai and Y. Yi, "Key frame extraction based on visual attention model," *Journal of Visual Communication and Image Representation*, vol. 23, no. 1, pp. 114–125, 2012.

[16] P. Pahalawatta, Z. Li, F. Zhai, and A. Katsaggelos, "Rate-distortion optimization for internet video summarization and transmission," in *MMSP'05*, 2005, pp. 1–4.

[17] E. K. Kang, S. J. Kim, and J. S. Choi, "Video retrieval based on key frame extraction in compressed domain," in *ICIP'99*, vol. 3, 1999, pp. 260–264.

[18] S. Baysal, M. Kurt, and P. Duygulu, "Recognizing human actions using key poses," in *ICPR'10*, 2010, pp. 1727–1730.

[19] Y. Ma and H. Zhang, "A model of motion attention for video skimming," in *ICIP'02*, vol. 1, 2002, pp. 129–132.

[20] M. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding," in *CAIVD'98*, 1998, pp. 61–70.

[21] M. Albanese, M. Fayzullin, A. Picariello, and V. Subrahmanian, "The priority curve algorithm for video summarization," *Information Systems*, vol. 31, no. 7, pp. 679–695, 2006.

[22] B. Chen, J. Wang, , and J. Wang, "A novel video summarization based on mining the story-structure and semantic relations among concept entities," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 295–312, 2009.

[23] J. Fan, A. Elmagarmid, X. Zhu, W. Aref, and L. Wu, "Classview: hierarchical video shot classification, indexing, and accessing," *IEEE Trans. Multimedia*, vol. 6, no. 1, pp. 70 – 86, feb. 2004.
[24] J. Sivic, F. Schaffalitzky, and A. Zisserman, "Object level grouping for video shots," *Int. J. Comput. Vision*, vol. 67, no. 2, pp. 189–210, 2006.
[25] C. Snoek, M. Worring, D. Koelma, and A. Smeulders, "A learned lexicon-driven paradigm for interactive video retrieval," *IEEE Trans. Multimedia*,, vol. 9, no. 2, pp. 280–292, 2007.
[26] A. Ekin, A. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Process.*, vol. 12, no. 7, pp. 796–807, 2003.
[27] K.-Y. Cheng, S.-J. Luo, B.-Y. Chen, and H.-H. Chu, "Smartplayer: user-centric video fast-forwarding," in *CHI'09*, 2009, pp. 789–798.
[28] J. Jiang, X.-P. Zhang, and A. C. Loui, "A content-based video fast-forward playback method using video time density function and rate distortion theory," in *ICME'11*, 2011, pp. 1–6.
[29] N. Petrovic, N. Jojic, and T. S. Huang, "Adaptive video fast forward," *Multimedia Tools Appl.*, vol. 26, no. 3, pp. 327–344, 2005.
[30] Z. Li, P. Ishwar, and J. Konrad, "Video condensation by ribbon carving," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2572–2583, 2009.
[31] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1971–1984, 2008.
[32] S. Feng, S. Z. Li, D. Yi, and Z. Lei, "Online content-aware video condensation," in *CVPR'12*, vol. 1, 2012, pp. 2082–2087.
[33] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, "Highlight sound effects detection in audio stream," in *ICME'03*, vol. 3, 2003, pp. 37–40.
[34] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, "Personalized abstraction of broadcasted american football video by highlight selection," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 575–586, 2004.
[35] B. Li, H. Pan, and I. Sezan, "A general framework for sports video summarization with its application to soccer," in *ICASSP'03*, vol. 3, 2003, pp. 169–172.
[36] F. Chen, C. De Vleeschouwer, H. Barrobes, J. Escalada, and D. Conejero, "Automatic summarization of audio-visual soccer feeds," in *ICME'10*, 2010, pp. 837–842.
[37] F. Chen and C. De Vleeschouwer, "Formulating team-sport video summarization as a resource allocation problem," *IEEE Trans. Circuits and Sys. Video Technol.*, vol. 21, no. 2, pp. 193–205, 2011.
[38] M. Höferlin, K. Kurzhals, B. Höferlin, G. Heidemann, and D. Weiskopf, "Evaluation of fast-forward video visualization," *IEEE Trans. Visualiz. and Computer Graphics*, vol. 18, no. 12, pp. 2095–2103, 2012.
[39] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, no. 9, pp. 1445–1453, 1988.
[40] A. Ortega, "Optimal bit allocation under multiple rate constraints," in *DCC'96*, 1996, pp. 349–358.
[41] H. Everett, "Generalized lagrange multiplier method for solving problems of optimum allocation of resources," *Operations Research*, vol. 11, no. 3, pp. 399–417, 1963.
[42] D. Helbing, P. Molnár, I. J. Farkas, and K. Bolay, "Self-organizing pedestrian movement," *Environment and Planning B-planning and Design*, vol. 28, pp. 361–383, 2001.
[43] F. Chen and A. Cavallaro, "Detection of group interactions by online association of trajectory data," in *ICASSP'13*, Paper No.2093, 2013.
[44] H. Pan, P. van Beek, and M. I. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *ICASSP'01*, vol. 3, 2001, pp. 1649–1652.
[45] I. Fernandez, F. Chen, F. Lavigne, X. Desurmont, and C. De Vleeschouwer, "Browsing sport content through an interactive H.264 streaming session," in *MMEDIA'10*, vol. 1, 2010, pp. 155–161.
[46] H. Duxans, X. Anguera, and D. Conejero, "Audio based soccer game summarization," in *BMSB'09*, 2009, pp. 1–6.
[47] J. Owens, "Tv sports production," *Focal Press*, 2007.
[48] A. G. Money and H. W. Agius, "Feasibility of personalized affective video summaries," in *Affect and Emotion in Human-Computer Interaction*, 2008, pp. 194–208.
[49] S. Zhao, H. Yao, X. Sun, X. Jiang, and P. Xu, "Flexible presentation of videos based on affective content analysis," in *MMM'13*, 2013, pp. 368–379.
[50] "Jaist multiview surveillance video dataset," 2012. [Online]. Available: http://www.jaist.ac.jp/%7echen-fan/multivision/jaistmvsdb.html
[51] "Behave: Computer-assisted prescreening of video streams for unusual activities," 2007. [Online]. Available: http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/
[52] D. Delannay, N. Danhier, and C. De Vleeschouwer, "Detection and recognition of sports(wo)men from multiple views," in *ICDSC'09*, 2009, pp. 1–7.
[53] F. Chen and C. De Vleeschouwer, "Partial motion trajectory grouping through rooted arborescence," in *ICIP'12*, Paper No.2511, 2012, pp. 1–4.
[54] "Tracking results of the surveillance dataset," 2012. [Online]. Available: http://www.jaist.ac.jp/project/prime-proj/results-tracking-tad.htm
[55] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, 2006.
[56] "Supplemental materials," 2012. [Online]. Available: http://www.jaist.ac.jp/project/prime-proj/tmm-supplementals.htm
[57] J. Nielsen, *Usability Engineering*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

Fan Chen is currently an assistant professor of Japan Advanced Institute of Science and Technology (JAIST) since 2010. He received his B.S., M.S. and Ph.D., from Nanjing University (China), Tohoku University (Japan), JAIST (Japan), in 2001, 2005 and 2008, respectively. He was supported by the Japanese Government MEXT Scholarship for foreign students, and received twice the awards for outstanding graduating students from both Nanjing Univ. (2001) and JAIST (2008), respectively. He was a COE Researcher for face recognition (JAIST, 2005-2007), a post-doctoral researcher in TELE, UCL, where he worked for the FP7 APIDIS European project (2008-2010), and an academic visitor to QMUL, UK (2012.02-04). His research interests are focused on statistical inference and optimization techniques related to computer vision, pattern recognition, and multimedia analysis.

Christophe De Vleeschouwer is a Senior Research Associate at the Belgian NSF, and an Associate Professor at UCL (ISPGroup). He was a senior research engineer with the IMEC Multimedia Information Compression Systems group (1999-2000), and contributed to projects with ERICSSON. He was also a post-doctoral Research Fellow at UC Berkeley (2001-2002) and EPFL (2004). His main interests concern video and image processing for content management, transmission and interpretation. He is enthusiastic about non-linear and sparse signal expansion techniques, ensemble of classifiers, multi-view video processing, and graph-based formalization of vision problems. He is the co-author of more than 30 journal papers or book chapters, and holds two patents. He serves as an Associate Editor for IEEE Transactions on Multimedia, has been a reviewer for most IEEE Transactions journals related to media and image processing, and has been a member of the (technical) program committee for several conferences. He contributed to MPEG bodies, and coordinated/participated to several European and Walloon Region projects (e.g. www.apidis.org). He is a co-founder of Keemotion (www.keemotion.com), using video analysis for autonomous content production.

Andrea Cavallaro is Professor of Multimedia Signal Processing and Director of the Centre for Intelligent Sensing at Queen Mary University of London, UK. He received his Ph.D. in Electrical Engineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, in 2002 and the Laurea (Summa cum Laude) in Electrical Engineering from the University of Trieste in 1996. He was a Research Fellow with British Telecommunications (BT) in 2004/2005 and was awarded the Royal Academy of Engineering teaching Prize in 2007; three student paper awards on target tracking and perceptually sensitive coding at IEEE ICASSP in 2005, 2007 and 2009; and the best paper award at IEEE AVSS 2009. Prof. Cavallaro is Area Editor for the IEEE Signal Processing Magazine; and Associate Editor for the IEEE Transactions on Image Processing. He is an elected member of the IEEE Signal Processing Society, Image, Video, and Multidimensional Signal Processing Technical Committee. Prof. Cavallaro was General Chair for IEEE/ACM ICDSC 2009, BMVC 2009, M2SFA2 2008, SSPE 2007, and IEEE AVSS 2007; and Technical Program chair of IEEE AVSS 2011; the European Signal Processing Conference (EUSIPCO 2008) and of WIAMIS 2010. He has published more than 130 journal and conference papers, and four books: Multi-camera networks (2009), Elsevier; Video tracking (2011), Wiley; Analysis, retrieval and delivery of multimedia content (2012), Springer and Intelligent multimedia surveillance (2013), Springer.