

Multi-view Video Summarization

ABSTRACT

Traditional video summarization algorithms focus on monocular videos. They are often limited in exploring the redundancy in multi-view videos and summarizing them into a concise representation. In this paper, we present for the first time a generic framework for multi-view video summarization. Essentially, we build a spatio-temporal shot graph and formulate the summarization problem as a graph labeling task. Specifically, we first partition the shot graph, and cluster event-centered shots with similar content by random walks. The summarization result is generated through solving a multi-objective optimization problem based upon shot importance robustly evaluated by a Gaussian entropy fusion scheme. The multi-objective optimization supports different summarization objectives, such as minimum summary duration and maximum information coverage. Moreover, multi-level summarization can be easily achieved by flexibly configuring parameters in optimization. Our experiments demonstrate the effectiveness of our method.

Keywords

Video summarization, Multi-view video, Gaussian entropy fusion, Random walks, Multi-objective optimization.

1. INTRODUCTION

With the development of computation, communication and storage infrastructures, multi-view video systems that simultaneously capture a group of videos and record the video content of the occurrence of events with considerable overlapping field of views (FOVs) across multiple cameras, have become more and more popular. In contrast to the quickly developing video collection and storage techniques, consuming these multi-view videos still remains a problem. For instance, browsing the large number of videos becomes a big challenge.

Video summarization techniques produce a condensed and succinct representation of video content, which facilitates

the browse and retrieval of the original video. There has been a rich literature on summarizing a long video into a concise representation, such as a video skim [] and a key-frame sequence [22]. These existing methods provide many effective solutions to help people browse a video clip. However, they focus on monocular videos. Directly applying these monocular video summarization methods to each component of a multi-view video group could lead to a redundant summarization result as each component has overlapping information with the others. Moreover, since multi-view videos often suffer from different lighting conditions in distinctive views, it is also non-trivial to equally evaluate the importance of shots of each view video and robustly merge each component into an integral video summarization.

In this paper, we present a generic framework for multi-view video summarization. We first parse the video from each view into shots. Content correlations of shots within each view as well as among multi-views intrinsically enable the spatio-temporal shot graph to be an optimal representation of multi-view videos, by viewing shots as graph nodes and mapping similarity of shots to edge weights. Specially, a Gaussian entropy fusion scheme is first proposed to evaluate importance of shots. We further use a random walk algorithm to partition this graph into event-centered shot clusters with similar video content. Finally, in a form of the graph labeling process, the summarization is generated through the multi-objective optimization.

The main contribution of this paper lies in that we present for the first time a multi-view video summarization method. It has the following notable features.

- We construct a spatio-temporal shot graph for the representation of multi-view video structure. Complicated content correlations among multi-view videos intrinsically make summarization difficult. The shot graph intuitively characterizes such correlations. Furthermore, we compute graph node value representing importance of shot by a Gaussian entropy fusion scheme. Such scheme, that obeys the inclusion-exclusion principle, can equally evaluate the shot importance in the presence of different lighting conditions and conspicuous noises, by emphasizing useful information while precluding redundancy among modeled features.
- We use random walks to cluster the event-centered clusters, and produce final summary by multi-objective optimization. Taking the shot graph as representation

makes the multi-view summarization tractable, since we can seek solution in the light of graph theory. We use random walks to first cluster those similar shots and take them as candidates for summarization. The shot clusters thus produced are event-centered that would facilitate shot category and retrieval. The multi-objective fuzzy optimization can be flexibly configured to meet various summarization requirements. In addition, multilevel summary can be easily achieved by setting different parameters.

The rest of this paper is organized as follows. We briefly review previous work in Section 2. In Section 3, we present a high-level overview of our framework. The two key components of our framework, spatio-temporal shot graph construction and multi-view summarization are presented in Section 4 and 5, respectively. We evaluate our method in Section 6, and conclude the paper in the last Section.

2. RELATED WORK

Most of previous work for video summarization basically involves two major steps. Videos are first clipped into basic temporal units. Afterwards, salient key-frames or important segments are selected according to user preference or different objectives.

For the first step, various algorithms have been proposed in the literature [14, 24]. Through analyzing slices extracted by partitioning video sequence and collecting temporal signature, Ngo *et al.* [14] proposed a video parsing approach. It is proved effective in detecting camera breaks like cuts, wipes, and dissolves. Xiang *et al.* [24] constructed a cumulative multi-event histogram over time to represent video content. An on-line segmentation algorithm named forward-backward relevance is developed to detect breaks in video content. Kraaij *et al.* [7] claimed that shot boundary detection has been considered a solved problem by the NIST TRECVID benchmark.

For the second step, many methods are proposed by utilizing varying kinds of mechanisms, such as simulated human perception [10, 13], graph analysis [25, 9, 13], etc. To reserve important shots in video summarization, saliency detection is often implemented by simulating human perception. In [6], Itti *et al.* devised a computationally efficient model of visual saliency for static image. Multiscale features are integrated into a saliency map, which is further refined by a dynamic neural network. Based on heuristic rules of human visual perception, Ma *et al.* [11] calculated visual-feature contrast as saliency. Then saliency detection is used to compute a normalized saliency value for each pixel. To evaluate visual saliency of video sequence, multi-modal features, such as motion vectors and audio frequencies should be taken into account. Ma *et al.* [10] created user attention curves in terms of visual, aural and linguistic features, respectively. These curves are fused further by linear or nonlinear schemes for video summarization. Integrating motion, contrast, special scenes, and statistical rhythm cues, You *et al.* [26] proposed a generic framework for human perception analysis. Using linear or priority based fusion approach, a perception curve is constructed for labeling three-level summarization, namely video key words, key frames, and dynamic segments.

Measuring shots' visual complexity and analyzing speech data, Sundaram *et al.* [21] generated audio-visual skims by constrained utility maximization that maximizes information content and coherence. Combining language and image understanding techniques, Smith *et al.* [20] extracted significant information, such as special objects, audio keywords, and relevant video structure for video skim. Nam *et al.* [12] detected emotional dialogues and violent scenes by investigating high variations of speech and video contents in spatial and temporal domain.

To extract hierarchically semantic or other high-level features, graph analysis is always employed in video summarization. Lu *et al.* [9] developed a graph optimization method that computes optimal video skim in each scene with dynamic programming. Ngo *et al.* [13] used temporal graph analysis to effectively capsule information for video structure and highlight. Through modeling the video evolution by temporal graph, their method can automatically detect scene changes and generate summaries. Sakarya *et al.* [18] proposed a graph-based multi-level temporal video segmentation method which partitions a weighted undirected graph into clusters in each level.

Moreover, some other approaches are employed to determine key-frames or important segments. Hanjalic *et al.* [5] divided video sequence into a number of clusters, and determined the optimal clustering by cluster-validity analysis. Each cluster is represented by one key-frame. DeMenthon *et al.* [2] regarded video sequence as a curve in high dimensional space. Then key-frames correspond to the control points that can provide a satisfying approximation to the original curve. After defining a metric for measuring missing frames and video summary distortions, Li *et al.* [8] solved video summarization using MINMAX optimization with viewing time, frame skip and bitrate constraints based on dynamic programming. By automatically learning users' understanding of the video content, Yu *et al.* [27] utilized previous viewers' browsing log to facilitate future viewers. They proposed a ShotRank metric to measure subjective interestingness and importance of video shot. Besides, there are methods using expectation maximization (EM) [15], singular value decomposition (SVD) [4], and etc. to obtain key-frames. For a thorough overview of related work on video summarization, please refer to [22].

The above methods contribute many effective solutions to monocular video summarization. They cannot, however, provide concise representation for multi-view summarization when applied directly to each view. Multi-view video coding (MVC) is an efficient approach to achieve the objective of compact representation. Using techniques such as motion estimation, disparity estimation, and etc., it removes information redundancy in spatial and temporal domain of videos. Nevertheless, MVC does not remove content redundancy so that it cannot facilitate browsing large volume of videos. This is exactly one objective of multi-view summarization. In this paper, some efforts are made to solve this problem.

3. OVERVIEW

Essentially, we construct a spatio-temporal shot graph to represent the multi-view videos, and to characterize multi-

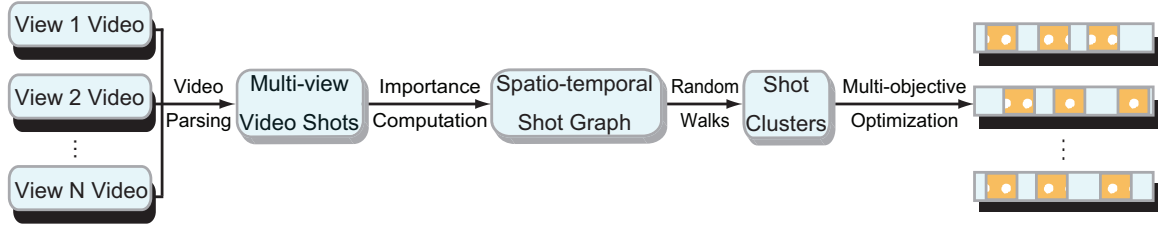


Figure 1: Overview of our multi-view video summarization method.

view correlations. Multi-view summarization is achieved through event-centered shot clustering via random walks and the multi-objective optimization. Spatio-temporal shot graph construction and the multi-view summarization are the two key components of our framework. To construct the shot graph, we first parse the input multi-view videos into content-consistent video shots. Dynamic and some important static shots are reserved as a result. For evaluating the importance of each shot, Gaussian entropy fusion model is developed to fuse together a set of intrinsic video features extracted. Afterwards, we construct the spatio-temporal shot graph by viewing shots as nodes and mapping correlations of shots in each view and across multi-views onto edge weights. To realize multi-view summarization on the graph, we employ random walks to cluster those event-centered similar shots. Using them as the anchor points, final summarization is produced by a multi-objective optimization model that supports various user requirements and the multi-level summarization. Fig. 1 illustrates the outline of our multi-view video summarization method.

In following, we elaborate on the two key components, i.e. spatio-temporal shot graph construction and the multi-view summarization.

4. SPATIO-TEMPORAL SHOT GRAPH

It is difficult to directly generate summarization, especially video skims from the multi-view videos. A common idea is to first parse the videos into shots. In this way, video summarization is transformed into a problem of selecting a set of representative shots. Obviously, the summarized shots should favor those interesting events occurred. Meanwhile, it is also invalid if the summarized shots are too trivial and discrete. To achieve this, content correlations among shots must be taken into account. For monoview videos previous summarization methods concern, each shot only correlates with its temporally adjacent shots. The correlations are simple, and easily modeled. However, for the multi-view videos, each shot correlates closely with not only its temporally adjacent shots in its own view, but also the spatially neighboring shots in other views. The correlations are non-linear and extremely complicated. To better explore such correlations, an appropriate representation of the multi-view structure which facilitates the multi-view summarization task is needed. We bring forward here a weighted spatio-temporal shot graph as the representation. The weights on edges just characterize complicated correlations of multi-view shots. By converting the multi-view shots into graph, more importantly, we can summarize the multi-view videos in the light of the well-studied graph theory and optimization, making summarization tractable.

4.1 Graph Construction

We first parse the multi-view videos into shots. In general, shot normally is an acceptable and adaptable granularity unit for describing the same event. We basically adopt the forward-backward relevance-based shot detection algorithm described in [24], while we further abandon those shots with lower activities. In particular, for every shot detected, we first compute the differential image sequence of adjacent frames. Each image can then be converted into a binary image by comparing the absolute value of each pixel against a threshold. We evaluate the activity of the shot through counting the total number of its non-zero pixels and comparing with a predefined activity threshold.

Parsing the multi-view videos into shots allows us to seek solution of summarization in a more compact shot space. Actually, each shot correlates with its temporal neighbors in its own view as well as the spatial neighbors in other views. Such characteristic makes the weighted graph an optimal representation of multi-view videos, by viewing shots as nodes and converting the correlations of shots into edge weights. Graph model has been used for monoview video summarization [18, 13, 9], in which the 2D graph is enough to represent the monoview video. In contrast, our graph here is a spatio-temporal shot graph. It has more complicated node connections due to the spatio-temporal correlations among multi-view shots (Fig. 2).

The multi-view videos are treated as a weighted undirected shot graph $G(V, E, W)$. Each node in V represents a shot resulting from video parsing. Its value R is the importance of shot calculated by the Gaussian entropy fusion model in the next subsection. The edge set E connects every pair of nodes if they are closely correlated. The correlation is just the edge weight measuring the visual and temporal similarity between shots S_i and S_j ,

$$W(S_i, S_j) = VisSim(S_i, S_j) * \frac{1}{\alpha_1 + \alpha_2 * d + \alpha_3 * d^2} \quad (1)$$

where $VisSim(S_i, S_j)$ is the visual similarity of shots S_i and S_j . For computational efficiency, we select three frames, namely the first, middle and last from S_i and S_j separately and calculate it according to their color histogram H_C and edge histogram H_E [1] distances,

$$VisSim(S_i, S_j) = w|H_C(S_i) - H_C(S_j)| + |H_E(S_i) - H_E(S_j)|. \quad (2)$$

Considering edge histogram weakens the influence of lighting difference across multi-view shots. w here is a weight that is empirically set to 0.5. d computes the temporal distance,

$$d = |t_i - t_j|, \quad (3)$$

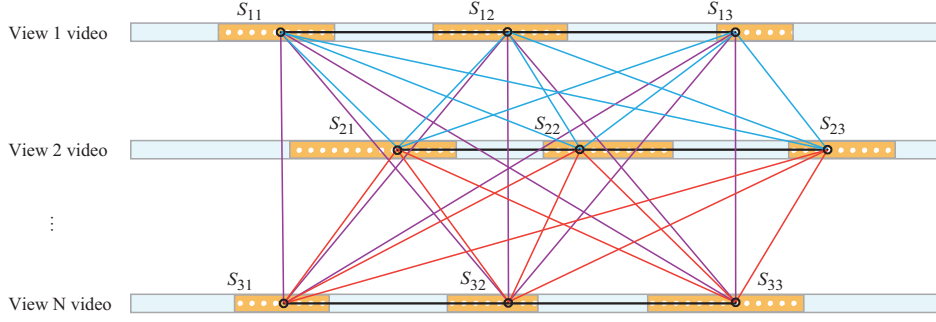


Figure 2: The spatio-temporal shot graph $G(V, E, W)$. Each node in G represents a shot, and its value R is the shot importance. Each edge connects a pair of nodes (shots) with correlation which is evaluated by shots' similarity. Without losing generality, only three shots in each view are given for illustration.

where t_i and t_j are the time of their middle frames. d is further integrated into a light attenuation function, in which the parameters α_1, α_2 and α_3 control the temporal similarity. They are set as 1, 0.01 and $1e-4$ respectively in our experiments.

Note that, for problem-specific domains, $VisSim$ could be modified to accommodate texture information, motion features etc. In addition, to further simplify the graph, W is set to zero if it is smaller than a predefined threshold.

By converting the multi-view videos into the spatio-temporal shot graph, the correlations of shots are naturally and intuitively reflected in the graph. Moreover, the graph nodes carry shot importance, which is necessary to create the concise and representative summary. We evaluate the importance by using a Gaussian entropy fusion model.

4.2 Importance Computation by Gaussian Entropy Fusion

To compute shot importance, previous monoview video summarization methods generally unite features with linear or non-linear fusion schemes. Such schemes, nevertheless, would not necessarily lead to the optimal performance for our multi-view videos in case the videos are contaminated by noises. This is especially true for those multi-view surveillance videos which often suffer from different lighting conditions across multiple views. Under such circumstance, we should robustly and equally evaluate the importance of the shots that may capture the same interesting event in different views and with different brightness. To account for this, we need to emphasize the portion of useful information in multi-view videos, and depress the influence of noises simultaneously. We first extract a set of intrinsic features from the videos. In most cases, the features extracted are definitely correlated with each other. Built upon such observation, a Gaussian entropy fusion model that essentially obeys the inclusion-exclusion principle is developed to glue them together.

Although feature extraction is beyond the state-of-the-art, low level visual features have been used successfully as an alternative. We now mainly take into account the visual features, such as color histogram feature, edge histogram feature, wavelet feature, motion vector feature, and the face detection feature. The features in other modalities, such as

textual and aural features used in previous video analysis methods [10] [26], however can also be integrated into our framework. Without losing generality, for shot S with n frames, suppose that overall M feature vector sets denoted by $\{F_i\}_{i=1}^M$ are extracted. Each feature F_i is expanded into a one column vector, and two arbitrary features F_i and F_j may have different dimensions.

The Gaussian entropy fusion model aims at emphasizing the implication of feature sets and minimizing noise influence. For shot S , we sum up the entropy values of all feature sets and subtract the entropy of their union from the sum,

$$R(S) = \sum_{i=1}^M H(F_i) - H(F_1, F_2, \dots, F_M), \quad (4)$$

where

$$H(F_i) = -p(F_i) \log(p(F_i)) \quad i = 1, \dots, M \quad (5)$$

and

$$\begin{aligned} H(F_1, F_2, \dots, F_M) \\ = -p(F_1, F_2, \dots, F_M) \log(p(F_1, F_2, \dots, F_M)). \end{aligned} \quad (6)$$

$F_i = (f_{i,1}, f_{i,2}, \dots, f_{i,n})_{i=1}^M$. $f_{i,j}$ is the i -th feature set for the j -th frame of shot S .

To estimate the probability of $p(F_i)$ and $p(F_1, F_2, \dots, F_M)$, a common idea is to approximate them with the Gaussian distribution,

$$p(F_i) \sim \mathcal{N}(\mathbf{0}, \Sigma^i) \quad (7)$$

$$p(F_1, F_2, \dots, F_M) \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad (8)$$

where Σ^i is the covariance matrix of $\{f_{i,j}\}_{j=1}^n$ ($i = 1, \dots, M$), and Σ is the one of $\{f_{i,j}\}_{i=1}^M$. F_1, F_2, \dots, F_M are normalized by,

$$f_{i,j}^* = \frac{f_{i,j} - \frac{1}{n} \sum_{j=1}^n f_{i,j}}{\sqrt{\frac{1}{n} \sum_{j=1}^n (f_{i,j} - \frac{1}{n} \sum_{j=1}^n f_{i,j})^2}}. \quad (9)$$

By virtual of non-linear time series analysis [17, 16, 19], the Gaussian entropy of shot S is finally expressed as,

$$R(S) = \frac{1}{2} \sum_{j=1}^n \log_2(\Sigma_{jj}) - \frac{1}{2} \log_2 |\Sigma| \quad (10)$$

where Σ_{jj} is the j -th element in the diagonal of matrix Σ . $|\Sigma| = \prod_{j=1}^n \lambda_j$. λ_j is the eigenvalue of Σ .

The entropy R is a measure of information encoded by shot S . We take it as the importance. An additive advantage of the Gaussian entropy fusion scheme is that it works well as long as the union of feature vector groups covers most useful information of multi-view videos. Therefore, instead of using all the feature sets listed, it would be sufficient if some well-defined feature sets are available.

5. MULTI-VIEW SUMMARIZATION

The spatio-temporal shot graph is an optimal representation of multi-view video structure, since it carries shot information and meanwhile reflects intuitively shot correlations. Owing to correlations of multi-view shots, the shot graph has complicated node connections. This makes the summarization task still challenging. We must generate those most representative graph nodes (shots) by taking into consideration the connections. It is unstraightforward. Fortunately, the well-studied graph theory enlightens us. Our basic observation here is that, with the shot graph, the multi-view video summarization can be formulated as a graph labeling problem. We accomplish this within two steps. We first cluster those event-centered similar shots, and pick out the candidates for summarization by random walks. Final summarization is produced by a multi-objective optimization process that is specifically devised to meet different user requirements.

5.1 Shot Clustering by Random Walks

To cluster similar shots, a few important shots are sampled. Using them as the anchor points, the shots that describe the same event are clustered together by random walks.

We adopt random walks in this step rather than other graph partition algorithms such as graph cut and normalized cut. Several reasons account for this. On the one hand, random walks has proven to be effective in handling large and complex graphs, even in the presence of conspicuous noises. It is thus suitable to our clustering task which needs to partition the spatio-temporal shot graph with complicated node connections. Graph cut however is prone to small cut and the noise influence. On the other hand, our graph partition is a K -way segmentation problem given sampled shots indicating seeds for candidate clusters. Random walks works well for it. The random walker starts from each unsampled node (shot) and determines for it the most preferable sampled

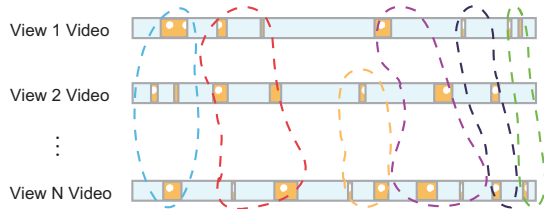


Figure 3: Graph partition by random walks. Shot clusters generated by random walks are enclosed by dashed circles.

shot's cluster. The final clusters thus obtained are actually event-centered since the weight on graph is defined in form of the visual and temporal similarity of shots. Although normalized cut can perform K -way segmentation, it however tends to produce an approximate solution that can be far from the global optimum. Furthermore, the computation of random walks is fast. The linear formulation of random walks allows it to be solved efficiently using a Conjugate Gradient algorithm.

Although a detailed description of random walks theory is beyond the scope of this paper, it essentially works as follows.

First, we partition the node set V into seeded nodes V_S and unseeded nodes V_U , satisfying that the value of each seed in V_S exceeds an entropy threshold.

We then define the combinatorial Laplacian matrix for graph as follows,

$$L_{ij} = \begin{cases} \sum_j W(S_i, S_j) & \text{if } i = j, \\ -W(S_i, S_j) & \text{if } S_i, S_j \text{ are linked by edge,} \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

L is a $N_n \times N_n$ dimensional sparse, symmetric and positive definite matrix, where N_n is the number of nodes in graph.

We further decompose L into blocks corresponding to nodes in V_U and V_S separately as,

$$L = \begin{bmatrix} L_S & B \\ B^T & L_U \end{bmatrix}. \quad (12)$$

For each unseeded seed, the final determination to which seeded cluster it belongs to is made by solving,

$$L_U X_U = -B^T X_S, \quad (13)$$

where X_U represents the probabilities that unseeded nodes belong to seeded nodes' clusters. X_S denotes the matrix that marks the cluster category of seeded nodes.

In the end, to favor important events of long duration, we filter out those trivial shot clusters with low entropy values. Furthermore, the two clusters whose similarity exceeds a given threshold are merged together. The shot clusters thus reserved are viewed as candidates for summarization in multi-objective optimization.

5.2 Multi-Objective Optimization

Users normally have various requirements over summarization, according to different kinds of application scenarios. Generally, we believe a good summary should achieve the following goals simultaneously. 1) Shot number minimization. The retrieval application of summary requires that a small number of shots should be generated. 2) Summary duration minimization. The minimum duration of summary would be of great help to video storage. 3) Information coverage maximization. To keep up enough information coverage, the sum of resulting shots' entropy value in each cluster must exceed a certain threshold. 4) Shot correlation maximization. It would be much better if shots in every resulting cluster strongly correlate with each other. This yields the most representative shots for the interesting event.

To meet the above goals, we devise a multi-objective optimization model to produce final summary. The optimization follows the complexity incompatibility principle [28]. We formulate summarization as a graph labeling problem. For the shot cluster C_S with n_s shots, the decision that whether or not the shots should be in the summary is denoted by $x = (x_1, x_2, \dots, x_{n_s})$, $\forall x \in X$. X is the 0/1 solution space in which $x_i = 1$ stands for reserved shot and 0 stands for unreserved one.

The multi-objective optimization function is given by,

$$\max\{-f_1(x), -f_2(x), f_3(x), f_4(x)\} \quad s.t. \quad \begin{cases} g(x) \leq D_{max} \\ h(x) \geq R_{min} \end{cases} \quad (14)$$

where $f_1(x) = \sum_{i=1}^{n_s} x_i$, $f_2(x) = \sum_{i=1}^{n_s} D_i * x_i$, $D_i > 0$,

$f_3(x) = \sum_{i=1}^{n_s} R_i * x_i$, and $f_4(x) = \frac{1}{2} \cdot \sum_{i,j=1, i \neq j}^{n_s} W(S_i, S_j) * x_i * x_j$.

D_i and R_i are duration and importance of shot i separately. $g(x)$ and $h(x)$ are defined in form of fuzzy set [3],

$$g(x) = \mu(f_2(x)), \quad h(x) = \mu(f_3(x))$$

with $\mu(f_i(x)) = [f_i(x) - \inf f_i(x)] / [\sup f_i(x) - \inf f_i(x)]$. D_{max} is the maximum allocated duration of one cluster. R_{min} is the minimum information entropy of C_S . They are defined as,

$$D_{max} = \lambda_1 \cdot D, \quad R_{min} = \lambda_2 \cdot R,$$

where D and R are the total duration of shots in C_S and the sum of importance values respectively. λ_1 and λ_2 are the parameters that control summary granularity. The two constraints mean that the total duration of shots in C_S after optimization should be less than D_{max} . Whereas the entropy should be greater than R_{min} . We will show in experiments, by flexibly configuring λ_1 and λ_2 , multi-level summarization can be easily achieved.

We further define the minimum function,

$$u(F(x)) = \min_{1 \leq i \leq 4} \{\eta_i \mu(f_i(x))\} \quad (15)$$

in which $F(x) = (\mu(f_1(x)), \mu(f_2(x)), \mu(f_3(x)), \mu(f_4(x)))^T$. $\eta_{i,i=1,\dots,4}$ are coefficients that control the weights of objective functions satisfying $\sum_{i=1}^4 \eta_i = 1$ and $\eta_i \geq 0$. They can be configured according to different user requirements.

By employing Max-Min method, the multi-objective optimization is transformed into the following 0-1 mixed integer programming problem:

$$x^* = \arg \max_{x \in X} u(F(x)) \quad s.t. \quad A \cdot F \leq \begin{pmatrix} D_{max} \\ -R_{min} \\ -u(F) \end{pmatrix}, \quad (16)$$

with $A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ -1 & -1 & -1 & -1 \end{pmatrix}$. x^* is the final optimization result to be solved.

The integer programming above can be solved by some classical algorithms. Nevertheless, since each cluster generally contains only a few shots, we use here brute-force search to find an approximate solution. It runs very fast for all our experiments.

6. EXPERIMENTS

The evaluation of video abstraction is a subjective work. "4C" Criterion is proposed to evaluate video summary [22]. Based on it, we raise four measurements: compressed ratio, information coverage, coherence, and multi-view redundancy.

Compressed ratio is computed by the following formulae:

$$CRatioF(S_f) = 1 - \frac{\gamma_{NKF}}{\gamma_{NF}}$$

where S_f is a given video series or video group sequences; $CRatioF(S_f)$ the compression rate calculated by the frame numbers of summary sequence and original ones; γ_{NKF} the number of frames in the summary; γ_{NF} the total number of frames in video sequence.

6.1 Experiments and Results Comparison

There is no standard dataset for testing multi-view abstraction we discussed above. For experimental purposes, we captured two multi-view surveillance video datasets named 2008-11-25 and 2009-01-08. As Table 1 show, 2008-11-25 is composed of three near-field office view videos. These videos are synchronous with good light conditions. Four semi-synchronous medium-field office videos make up the dataset of 2009-01-08. Further, in order to test the robustness of Gaussian entropy fusion model, different lighting conditions are provided. Please visit our website to download the datasets and our demos.

Normally, it is hard to give sound evaluations of our methods owing to the first time of paying attention to multi-view video abstraction. However, two aspects of comparisons are needed. On the one hand, we compare our results with the original videos listed in Table 1. For more impressive comparison, please visit our website. On the other hand, we compare our work with two mono-view video abstraction methods, namely Lu's spatio-temporal graph optimization [9] and Ma's user attention model [10].

The evaluations of information coverage, coherence and multi-view redundancy are based on user test scores. In our tests, visual and motion features are taken into account. So we adjust fusion schemes correspondingly in [10]. To equally evaluate these three methods, we partition our datasets with our video parsing algorithm in Section ??, and discard some static frames that contain nothing but the pictures of the test environment. For convenient evaluations, a maximum shot duration is set to 1.5%, 3%, 4%, while the information coverage is set to 50%, 60%, 70%. We generated 3 groups of video skim from the test video datasets. The 20 volunteers were invited to assess the video summaries by the following methods [26]. For comparison, we fixed the length of summaries and listed five score levels of 100%, 75%, 50%, 25% and 0, respectively. The volunteers gave their judgement from the former criteria. And they were required to rank only one score for each dataset from the viewpoint of multi-view redundancy. We averaged the values and compared them in Table 2 and 3. Table 2 contains the data of the

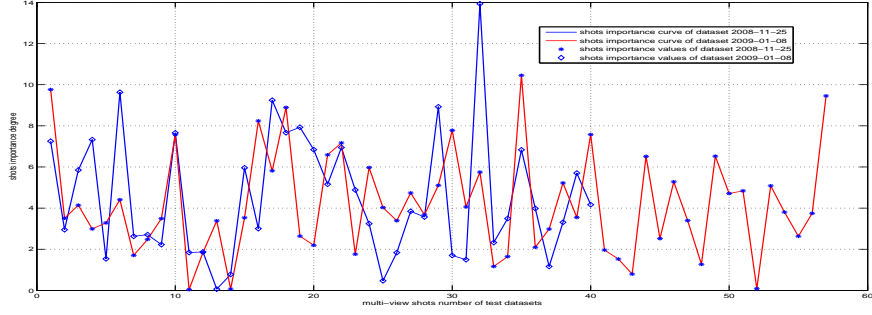


Figure 4: Importance curve of multi-view’s preserved shots in the two testing video datasets. All the preserved shots within one dataset are orderly arranged. To get the details of these shots in each view videos, please refer Table 1.

reduction degree of multi-view redundancy. Informativeness and coherence are calculated and listed in Table 3.

6.2 Evaluations and Some Analysis of Experiments

According to the evaluations, volunteers do not always give our method higher scores than that of others from the viewpoints of information coverage and coherence in single view video’s evaluations. However, when it comes to the reduction degree of multi-view redundancy, they rank our methods relatively high scores. There are some analysis of our experiments.

Admittedly, some important shots will be the candidates for all these three methods. Yet, they make great effects to almost all the view videos. Similar important content would be grasped by all view’s video skim. The proposed framework automatically restricts the influences of these important shots by partitioning several clusters and decreases the probability of repeated video skim to important shots among all of the views. Further, it is available to provide an exact semantic tag to almost every shot selected, since one cluster mainly describes one event. Furthermore, due to confining the influences of important events, some abnormal shots are also captured in our results, while they are likely to be neglected for low influences in the other two methods. For example, in our test dataset 2009-01-08 the sudden change of laptop’s screen can be captured in our method. Therefore, our method achieves the goals of providing a stable multi-view video abstraction method, effectively compressing all the multi-view video data and extracting more useful information.

The process of selecting important shots is also deserved to explain and make some comparison. In one view video, user attention model is inclined to pick up shots from the attention curve, while spatio-temporal graph optimization would prefer to choose longer shots based on the scene entropy formulae they proposed. In our approach, a user-defined process is presented for each event-centered cluster. Based on the four goals listed above, it facilitates satisfying different requirements of our viewers. This is another reason why our evaluations are relatively higher.

Table 2: Multi-view redundancy comparison within the three video abstraction algorithms

Video Abstraction Methods	Multi-view	Lu’s	Ma’s
Redundancy of 1,2,3	0.69	0.60	0.59
Redundancy of 4,5,6,7	0.80	0.64	0.73

6.3 Automatical semantic category and shot retrieval

Besides video summarization, automatic semantic category and shot retrieval are two of the most significant applications in our graph model. Since our model uses random walks as an event-centered clustering process, the shots within one cluster should describe the same or similar events. It is available to automatically define the semantic category of one cluster according to the semantic described seed shot. Admittedly, our methods do not bridge the semantic gap between low-level features and high level features. However, it provides a feasible method to diffuse the semantic description from seed shot.

Another interesting application is shot retrieval. Normally, we provide two methods for this purpose. The first one is to directly employ the concept of cluster obtained in last section. And if one input a shot for retrieval purpose, all the shot in the same cluster is returned to users. In our experiment, it is employed and Fig. ?? illustrates this idea. The second one is aiming to repeat the process of random walks by using the retrieval shot as the seed node. If the probability of some shot exceeds a threshold predefined, it is also a result for retrieval.

7. CONCLUSION AND FUTURE WORK

In this paper, we proposed a multi-view video abstraction framework that includes three major models. Video parsing decomposes the video into an adaptive granularity of video units. Feature vectors are extracted in each video unit and we robustly fuse them by Gaussian entropy fusion model for shots importance degree. Finally, a graph model for multi-view videos is constructed. After event-centered clustering and graph segmentation by random walks, multi-objective fuzzy optimization is utilized to balance all kinds of requirements for video summary. Experiments show that our methods have an excellent video compression ability while re-

Table 1: Description of testing videos and a summary result. SShots is Summary shots results. TCR means total compression rate: $CRatioF(summaryshots) = 1 - \frac{summaryShotsFrames}{originalVideosFrames}$.

No.	Video Name	Shots	SShots	Length	Slength	TCR
1	2008-11-25-view1 video	95	8	35:10	00:15	99.03%
2	2008-11-25-view2 video	73	3	35:03	00:17	
3	2008-11-25-view3 video	82	4	35:20	00:14	
4	2009-01-08-view1 video	67	7	11:16	00:21	95.29%
5	2009-01-08-view2 video	54	7	08:43	00:36	
6	2009-01-08-view3 video	61	4	11:22	00:31	
7	2009-01-08-view4 video	72	10	14:58	00:33	
Total	—	504	43	151:22	03:02	—

Table 3: Evaluation of the three video abstraction algorithms from the viewpoint of information coverage and coherence. Single view video evaluation and multi-view evaluation (total column value) are combined.

Predefined Length	1.5%			3%			4%		
Video Abstraction Methods	Multi-view	Lu's	Ma's	Multi-view	Lu's	Ma's	Multi-view	Lu's	Ma's
1	0.68,0.73	0.53,0.49	0.60,0.56	0.84,0.73	0.68,0.73	0.61,0.70	0.90,0.76	0.30,0.43	0.68,0.58
2	0.48,0.80	0.67,0.71	0.56,0.75	0.68,0.71	0.68,0.70	0.58,0.66	0.79,0.76	0.79,0.78	0.54,0.69
3	0.64,0.61	0.43,0.51	0.69,0.66	0.66,0.78	0.69,0.71	0.71,0.63	0.64,0.66	0.65,0.68	0.76,0.68
4	0.58,0.64	0.51,0.63	0.61,0.73	0.58,0.74	0.65,0.65	0.70,0.65	0.70,0.60	0.66,0.70	0.51,0.68
5	0.45,0.64	0.33,0.55	0.53,0.49	0.38,0.61	0.51,0.56	0.71,0.53	0.44,0.55	0.60,0.55	0.44,0.55
6	0.39,0.69	0.59,0.59	0.54,0.63	0.70,0.73	0.44,0.66	0.70,0.64	0.66,0.71	0.71,0.66	0.54,0.61
7	0.75,0.75	0.73,0.71	0.83,0.73	0.65,0.70	0.75,0.60	0.80,0.81	0.93,0.80	0.75,0.66	0.55,0.65

tain reasonable information coverage. Compared with single view video abstraction methods, our methods use the summarized results of each view video compensated for each other and have a global improvement on the measures of “4C” criteria over simple plus many single view algorithm’s summarized results.

There are also some limitations and future works. First, our works are mainly confined within the analysis of near-field and medium-field video [23]. Because the far-field’s visual changes are difficult to calculate and compare. Therefore, machine learning technology is needed such as support vector machine (SVM), hidden markov model (HMM) and neural network (NN). Second, the compression ratio has tight relation with the video content. For example, with saving the same information coverage of original videos, the videos compression ratio of a busy workday in bank office should be bigger than that of a holiday. Third, in our test we only summarize surveillance multi-view videos in office environment. However, much more work should be done in order to explore the applications in various other multi-view video types, for example, sport videos.

8. REFERENCES

- [1] G. Ciocca and R. Schettini. An innovative algorithm for key frame extraction in video summarization. *Journal of Real-Time Image Processing*, 1:66–88, Sep 2006.
- [2] D. Dementhon and D. Doermann. Video summarization by curve simplification. In *ACM Multimedia*, pages 211–218, 1998.
- [3] L. Dengfeng and C. Shouyu. A fuzzy programming approach to fuzzy linear fractional programming with fuzzy coefficients. *Fuzzy Mathematics*, 4(4):829–833, 1996.
- [4] Y. Gong and X. Liu. Video summarization and retrieval using singular value decomposition. *Multimedia Systems*, 9:157–168, Aug 2003.
- [5] A. Hanjalic and H. Zhang. An integrated scheme for automated video abstraction based on unsupervised cluster-validaty analysis. *IEEE Trans. on Circuits and Systems for Video Technology*, 8:1280–1289, 1999.
- [6] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, March 2001.
- [7] Kraaij, Wessel, Smeaton, F. Alan, and O. Paul. Trecvid 2004 - an overview. In *Text Retrieval Conf. TRECVID Workshop*, Nov 2004.

- [8] Z. Li, G. M. Schuster, and A. K. Katsaggelos. Minmax optimal video summarization. *IEEE Trans. on Circuits and Systems for Video Technology*, 15:1245–1256, Sep 2005.
- [9] S. Lu, K. Irwin, and M. R. Lyu. Video summarization by video structure analysis and graph optimization. In *Proc. IEEE Int. Conf. Multimedia and expo*, Jun 2004.
- [10] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Trans. on Multimedia*, 7(5):907–919, Oct 2005.
- [11] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *ACM Multimedia*, pages 374–381, 2003.
- [12] J. Nam and A. H. Tewfik. Dynamic video summarization and visualization. In *ACM Multimedia*, pages 53–56. ACM, 1999.
- [13] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang. video summarization and scene detection by graph modeling. *IEEE Trans. on Circuits and Systems for Video Technology*, 15:296–305, Feb 2005.
- [14] C.-W. Ngo, T.-C. Pong, and R. T. Chin. Video partitioning by temporal slice coherency. *IEEE Trans. on Circuits and Systems for Video Technology*, 11:941–953, Aug 2001.
- [15] X. Orriols and X. Binefa. An em algorithm for video summarization, generative model approach. In *Proc. IEEE ICCV*, pages 335–342, Aug 2002.
- [16] M. Palus. Testing for nonlinearity using redundancies: Quantitative and qualitative aspects. *Physica D*, 80:186–205, 1995.
- [17] M. Palus, V. Albrecht, and I. Dvorak. Information theoretic test for nonlinearity in time series. *Physics Letters A*, 175:203–209, 1993.
- [18] U. Sakarya and Z. Telatar. Graph-based multilevel temporal video segmentation. *Multimedia Systems*, 14(5):277–290, Nov 2008.
- [19] R. H. Shumway and D. S. Stoffer. *Time Series Analysis And Its Applications*. Springer Verlag, 2000.
- [20] M. A. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding techniques. In *Proc. IEEE CVPR*, pages 775–781, 1999.
- [21] H. Sundaram, L. Xie, and S. fu Chang. A utility framework for the automatic generation of audio-visual skims. In *ACM Multimedia*, pages 189–198, 2002.
- [22] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1):3, Feb 2007.
- [23] P. Turaga, R. Chellapa, V. S. Subramanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Trans. on Circuits and Systems for Video Technology*, 18(11):1473–1487, Nov 2008.
- [24] T. Xiang and S. Gong. Activity based video content trajectory representation and segmentation. *BMVC*, 2004.
- [25] M. M. Yeung and B.-L. Yeo. Video visualization for compact presentation and fast browsing of pictorial content. *IEEE Trans. on Circuits and Systems for Video Technology*, 7:771–785, Oct 1997.
- [26] J. You, G. Liu, L. Sun, and H. Li. A multiple visual models based perceptive analysis framework for multilevel video summarization. *IEEE Trans. on Circuits and Systems for Video Technology*, 17(3):273–285, 2007.
- [27] B. Yu, W.-Y. Ma, K. Nahrstedt, and H.-J. Zhang. Video summarization based on user log enhanced link analysis. In *ACM Multimedia*, pages 382–391, New York, NY, USA, 2003.
- [28] L. A. Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, Sep 1965.