Video Summary Quality Evaluation Based on 4C Assessment and User Interaction

Tongwei Ren, Yan Liu, and Gangshan Wu

Abstract As video summarization techniques have attracted increasing attention for efficient multimedia data management, quality evaluation of video summary is required. To address the lack of automatic evaluation techniques, this chapter proposes a novel full-reference evaluation framework to assess the quality of the video summary according to various user requirements. First, the reference video summary and the candidate video summary are decomposed into two sequences of Summary Units (SUs), and the SUs in these two sequences are matched by frame alignment. Then, a similarity-based assessment algorithm is proposed to automatically provide comprehensive human-like evaluation results of the candidate video summary quality from the perspective of Coverage, Conciseness, Coherence, and Context (4C), respectively. Considering the evaluation, criteria of video summary quality are usually application-dependent, the incremental user interaction is utilized to gather the user requirements of video summary quality, and the required evaluation results are transformed from the 4C assessment scores. The proposed framework is experimented on a standard dataset of TRECVID 2007 and shows a good performance in automatic video summary evaluation.

1 Introduction

The exponential growth of multimedia data and the wide application of multimedia technology have led to the significant need for efficient multimedia data man-

T. Ren (✉) · G. Wu
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
e-mail: rentw@graphics.nju.edu.cn

G. Wu
e-mail: gswu@nju.edu.cn

Y. Liu
Department of Computing, The Hong Kong Polytechnic University, Hong Kong, Hong Kong
e-mail: csyliu@comp.polyu.nju.hk

L. Shao et al. (eds.), *Multimedia Interaction and Intelligent User Interfaces*,
Advances in Pattern Recognition,
DOI 10.1007/978-1-84996-507-1_10,

agement [19]. Video summarization provides a means to manage video collections more efficiently by generating a concise statement, called a summary, in such a way that the user can understand the content of the video file(s) by merely viewing the summary [12]. A good video summary epitomizes the essentials of the original video in the form of storyboard (a collection of still images) [30] or video skim (a much shorter video clip) [17]. An informative and concise video summary enables efficient access to the voluminous, redundant, and unstructured video collections [5].

Although video summarization has received more and more attention, a systematic evaluation framework for video summarization is still unavailable [29]. Currently, the quality of the video summary is mainly assessed by human individuals [1, 14, 28], which is seriously influenced by human factors. Moreover, this kind of manual evaluation has high labor cost and time cost [23]. The missing of the automatic evaluation in video summarization also results in the problem that each work on video summarization may demonstrate its performance using its own evaluation method and often be short of the performance comparison with different techniques [29].

Due to the limitation of manual evaluation for video summary, automatic evaluation techniques providing the human-like assessment are highly demanded [15]. Some work has been done to evaluate the quality of the video summary by automatically calculating the inclusion and redundancy based on predefined ground truth [4, 8, 27, 32]. However, the uniform framework with comprehensive consideration for automatic evaluation is still missing. For example, the correct order of the content is very important for a good video summary, but this criterion and its interaction with other criteria have not been fully explored by current work. Moreover, the existing automatic evaluation techniques only provide the evaluation results according to their defined criteria respectively. They cannot satisfy the various user requirements of video summary quality in different applications.

To address the problem of current work on automatic evaluation for video summary, we propose a uniform framework providing automatic video summary quality evaluation according to various user requirements. The framework focuses on full-reference quality evaluation for video summary, meaning that the candidate video summary is evaluated based on the comparison with a predefined reference video summary. Full-reference quality evaluation is initially defined by Wang et al. to evaluate the quality loss of the image after some processing via comparing with a complete perfect reference [31]. Relatively, there exist nonreference quality evaluation [25] and reduced-reference quality evaluation [10] when the reference is not or only partially available. Considering the users may have more ambiguous perception of the perfect video summaries than in other applications, e.g., image compression, we utilize one or several defined reference video summaries to represent the perfect summaries and eliminate the inconsistence in evaluation. Furthermore, to satisfy various user requirements of video summary quality in different applications, we divide the whole evaluation procedure into two steps. We first generate a requirement-independent intermediate evaluation results by assessing the video summary quality according to a general criteria and then transform the intermediate

evaluation results to the final ones to satisfy the user requirements. In this chapter, we utilize the 4C criteria in [6] as the intermediate evaluation criteria. It provides a comprehensive description of video summary quality, including the aspects of information representation, such as coverage and conciseness, and the aspects of user perception, such as coherence and context. The existing human-like evaluation criteria can be mainly derived from the criterion or combinations of the criteria in these four aspects. In the evaluation framework, we propose several novel methods to calculate the scores on these criteria automatically. With the 4C assessment results, we use the incremental user interaction to gather the necessary information of user requirements and generate the required evaluation results by automatically transforming the 4C assessment scores.

The chapter is organized as follows. Section 2 introduces current quality evaluation methods for video summary. Section 3 proposes a novel framework of video summary quality evaluation and some initial processing algorithms, such as summary unit generation and matching. Section 4 provides the automatic 4C assessment algorithm for providing comprehensive intermediate evaluation results. Section 5 presents the transformation between the 4C assessment scores and the required evaluation results using user-interaction-based automatic transformation. Section 6 shows the performance of the proposed framework and techniques by experimenting on the standard datasets. The chapter is closed with conclusion and further work.

2 Related Work

Referencing the classification of text summarization assessment [2], quality evaluation of video summary can be classified into two categories, intrinsic evaluation and extrinsic evaluation. The former tests the summaries by themselves, while the latter tests the summaries based on how they interact with the completion of some other tasks. In this chapter, we use intrinsic evaluation to assess the video summary quality.

Based on the difference of human's interaction, current quality evaluation methods for video summary can be further categorized to manual evaluation and automatic evaluation [29]. Manual evaluation mainly involves independent users judging the quality of the generated video summaries and calculates the cognitive value based on psychological metrics [8]. The direct and the most widely used manual evaluation is asking the different persons to grade the summary individually and calculate the mean opinion score (MOS) as the quality score of the summary [9]. But only using the overall score is too rough in evaluation. So different evaluation criteria are proposed to define the desirable characters for a good summary. A typical set of evaluation criteria was proposed by He et al. [6], who provided the 4C criteria for an ideal video summary:

– *Coverage*: the set of segments selected for the summary should cover all the "key" points.
– *Conciseness*: any segment selected for the summary should contain only necessary information.

– *Coherence*: the flow between the segments in the summary should be natural and fluid.
– *Context*: the segments selected and their sequencing should be such that prior segments establish appropriate context.

Existing work of manual evaluation can be mainly recapitulated by the criteria or combinations of the criteria under these 4C criteria. For example, in the task of rushes summarization for TRECVID 2007 [22], the criterion of ground-truth inclusion actually can be considered as one way to measure the coverage of the summary.

Although manual evaluation is probably the most useful and realistic form of video summary evaluation [29], it suffers from several problems. First, manual evaluation is seriously affected by human factors [22]. Illustrated using TRECVID 2007 rushes summarization task, one evaluator is asked to evaluate four hundred and thirty-two video clips from eighteen rushes files. Moreover, for each rushes file, the evaluator should assess twenty-four very similar summaries. Consequently, it is so difficult to guarantee that the evaluator can keep the consistent scoring criterion throughout the evaluation, although he may intend to [22]. The human factors of manual evaluation can be removed or partially removed by some statistical techniques based on large dataset experiments. Unfortunately, it leads to high labor cost and huge time consumption [29]. For these reasons, the large user-set study is not widely employed [8]. Even for the TRECVID 2007, each video summary is only evaluated by three persons, which is far from what is required by statistical sufficiency. In addition, the invested labor and time in the user study for evaluating one algorithm is not reusable for another algorithm; all the effort has to be repeated each time when an algorithm has been changed or a new algorithm has been developed [7].

Due to the limitation of manual evaluation, the automatic evaluation techniques for video summary are highly demanded. Currently, automatic evaluation techniques can be classified into two categories. One category focuses on assessing the objective criteria, such as the length of the summary [22], while another category works on providing the human-like assessment by quantitative analysis of multimedia content. To map human's judgment, most automatic evaluation methods manually define a set of ground-truth or/and keyframes. Silva et al. [27] and Yahiaoui et al. [32] calculate the coverage of video summary by using the total keyframe number in summary or keyframe number in average keyframe set in place of the ground truth inclusion. Huang et al. [8] calculate precision, recall, and redundancy rate by matching the predefined ground truths in order to evaluate the content coverage and redundancy of video summary. Dumont et al. [4] use machine learning methods to train the automatic assessors on the manually generated ground truth and evaluate the ground truth inclusion of video summary by the assessors. Unfortunately, these methods only provide the evaluation of video summary quality in one or several aspects. Some important factors influencing video summary quality, such as the order of video content, are ignored in current work. Moreover, the existing automatic methods can only provide the evaluation results according to their defined criteria. The users cannot obtain the quality evaluation of video summary according to the

requirements outside these criteria. Till now, a uniform framework with comprehensive considerations of automatic evaluation for video summary is unavailable yet.

3 Uniform Framework for Video Summary Quality Evaluation

Figure 1 shows the framework of full-reference quality evaluation system for video summary. The reference video summary is assumed to be the only perfect abstraction of the original video file, which can be automatically or manually generated by any approaches or tools. This means that a candidate video summary will obtain a full mark in any criterion of evaluation if and only if the candidate video summary is the same with the reference one in this criterion. If there exist more than one reference summary, the evaluation is carried out on each reference video, and the best evaluation result is chosen as the final result. In this way, full-reference video summary quality evaluation is formalized to the problem of pair-wise video sequence comparison for evaluation purpose.

Although many techniques have been proposed to compare the similarity of the video sequences [13, 26], none of them have been successfully applied to video summary quality evaluation because of the different targets of the tasks. Most existing works of video sequence comparison are designed for video retrieval and classification [19], so they focus on providing qualitative results, for example, relevance or irrelevance for video retrieval. In other words, the target of these algorithms is to capture the main content while keeping insensitive to the details. But for video summary evaluation, the main content of the candidate summaries are almost identical. The difference of certain kinds of details often represents the difference of the quality. Therefore, these existing video sequence comparison techniques are not directly applicable to video summary evaluation. In this chapter, we address the problem by aligning the video summaries and compare the video summaries based on the alignment result. Considering that the video content may be represented with incorrect order in the candidate summary, we first decompose the reference summary and the candidate summary into a set of Summary Units (SUs) respectively and then apply frame alignment algorithm in matching the SUs from the two summaries.

Based on the SU matching result, we compare the reference summary and the candidate summary for quality evaluation. To provide a flexible evaluation mecha-

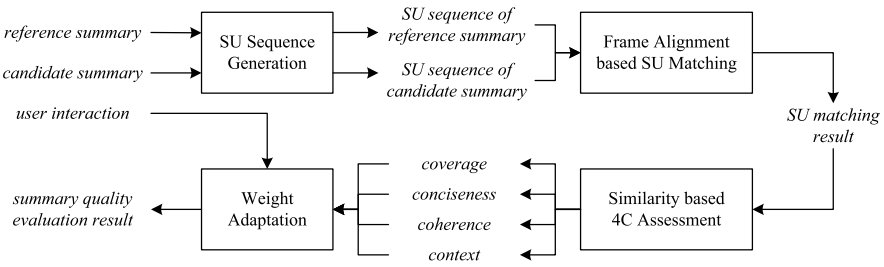

Fig. 1 Uniform framework for quality evaluation of video summary

nism satisfying various user requirements, we divide the following evaluation into two steps that generate the requirement-independent intermediate evaluation results and transform the intermediate evaluation results to the final results satisfying user requirements respectively. We first calculate the quality scores of the candidate summary individually in four aspects: coverage, conciseness, coherence, and context, which are derived from the 4C criteria in [6] and treated as the intermediate evaluation results. Then, we utilize the incremental user interaction to gather user requirements of video summary quality. The users are asked to manually evaluate some training data according to their required criteria. Based on the user interaction, the transformation model from the 4C assessment scores to the required evaluation results is generated. For the different candidate summaries evaluated by the same criteria, only once user interaction and transformation model generation are needed. Finally, the evaluation results of the candidate summary quality are generated by automatic transformation.

3.1 Summary Unit Sequence Generation

Simply speaking, summary unit is defined as the component to compose a video summary. It can be a video scene, shot, subshot, and even a frame for different video files and different summarization targets. Definitely, if the spatial separability is permitted, SU can be a special region of the frame or an object, and if the spatial-temporal separability is permitted, SU can be defined as a trajectory. Moreover, SU also can be a data package of synchronized or unsynchronized video, audio, and close caption. Due to the page limitation, we only consider the temporal separability of the video file for video summary quality evaluation, e.g., subshot is used as SU in this chapter. Thus a video summary can be described as an SU sequence with the appropriate order.

Considering a video summary S with N SUs, it can be represented using the SU sequence $S = \{SU_1, SU_2, \ldots\}_N$. Hence, the reference summary and the candidate summary can be represented as follows:

$$\begin{aligned} RS &= \{SU_{R_1}, SU_{R_2}, \ldots\}_{N_R}, \\ CS &= \{SU_{C_1}, SU_{C_2}, \ldots\}_{N_C}, \end{aligned} \tag{1}$$

where RS and CS denote the reference summary and the candidate summary, and N_R and N_C are the SU numbers of RS and CS, respectively. The following evaluation is based on the comparison of these two SU sequences. In this chapter, we generate the SU sequences using the twin-comparison algorithm in [33].

3.2 Frame Alignment-Based Summary Unit Matching

After SU sequence generation, we build the comparison between the reference summary and the candidate summary on the basis of SU matching. This means that we

check each SU in the candidate summary by looking for the most similar SU in the reference summary and compare the reference summary and the candidate summary based on the SU matching result. Various algorithms are available for subshots matching [4, 20]. Considering the requirement of matching accuracy, we treat SU as a time-order frame sequence and match SUs by aligning the corresponding frame sequences with the Needleman–Wunsch algorithm [21]. The frame alignment-based SU matching method can provide more accurate matching result than clustering-based methods for it can distinguish the detail differences between two adjacent SUs with similar content.

We represent SU_{R_i} in the reference summary as a frame sequence $\{f_1^{R_i}, f_2^{R_i}, \ldots\}_{m_i}$ and SU_{C_j} in the candidate summary as a frame sequence $\{f_1^{C_j}, f_2^{C_j}, \ldots\}_{n_j}$, where m and n are the frame numbers of SU_{R_i} and SU_{C_j}, respectively. Then, we use the Needleman–Wunsch algorithm to achieve the optimal matching of SU_{R_i} and SU_{C_j}. The Needleman–Wunsch algorithm utilizes dynamic programming in alignment, and the objective function is defined as follows:

$$\begin{aligned} s_{p1} &= \chi\left(f_p^{R_i}, f_1^{C_j}\right), \\ s_{1q} &= \chi\left(f_1^{R_i}, f_q^{C_j}\right), \\ s_{pq} &= \max\left(s_{p(q-1)}, s_{(p-1)q}, s_{(p-1)(q-1)} + \chi\left(f_p^{R_i}, f_q^{C_j}\right)\right), \end{aligned} \tag{2}$$

where $\chi(f_p^{R_i}, f_q^{C_j})$ is a function to denote whether $f_p^{R_i}$ and $f_q^{C_j}$ can be matched.

In our previous work [24], we utilize the similarity on local HSV color histogram to judge whether two frames can be matched. Though local HSV color histogram has good performance in video content similarity measurement, it cannot effectively distinguish the video frames with similar content but different details. It may lead to the inaccuracy in SU matching and influence the further quality evaluation. Therefore, we use Scale-Invariant Feature Transform (SIFT) [16] instead of local HSV color histogram in this chapter, which is effective in distinguishing different visual content and widely used in near-duplicate video detection [3, 11]. For each frame $f_p^{R_i}$ in SU_{R_i} and each frame $f_q^{C_j}$ in SU_{C_j}, we detect the keypoints in them with Hessian Affine detector and match the keypoints in the two frames by calculating their local gradient histogram distance. If the local gradient histogram distance of two keypoints is smaller than a predefined threshold (usually 0.3), the two keypoints are matched; otherwise, they are not matched. Note here that in order the matched frames in alignment to be highly similar, each keypoint kp_x in $f_p^{R_i}$ (or kp_y in $f_q^{C_j}$) is only required to look for its matched keypoint within the 16×16 neighboring region around the corresponding position in $f_q^{C_j}$ (or $f_p^{R_i}$). This constraint can well reduce the computational cost and avoid the keypoint mismatching. If no such keypoint exists, the contribution of keypoint kp_x (or kp_y) in frame matching is set to 0; otherwise, the contribution of the two keypoints is both set to 1. Each keypoint is only allowed matching one keypoint, and the match value of $f_p^{R_i}$ and $f_q^{C_j}$ is calculated

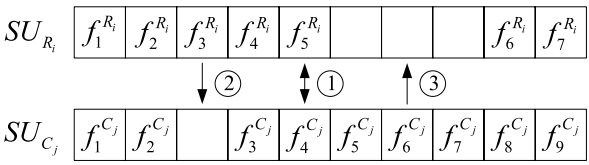

Fig. 2 Aligned frame sequences of two SUs in the reference summary and the candidate summary

as follows:

$$Mat\left(f_p^{R_i}, f_q^{C_j}\right) = \frac{1}{N_p^{R_i}} \sum_{kp_x \in f_p^{R_i}} \varphi(kp_x) + \frac{1}{N_q^{C_j}} \sum_{kp_y \in f_q^{C_j}} \varphi(kp_y), \tag{3}$$

where $N_p^{R_i}$ and $N_q^{C_j}$ are the numbers of keypoints in frames $f_p^{R_i}$ and $f_q^{C_j}$, respectively, $\varphi(kp_x)$ and $\varphi(kp_y)$ denote the contributions of keypoints kp_x and kp_y in frame matching, respectively, and $Mat(f_p^{R_i}, f_q^{C_j})$ denotes the matching value of $f_p^{R_i}$ and $f_q^{C_j}$. If the matching value of $f_p^{R_i}$ and $f_q^{C_j}$ is larger than a predefined threshold thr_{fm} ($thr_{fm} = 0.6$ in our experiments), we consider the two frames as matched, i.e., $\chi(f_p^{R_i}, f_q^{C_j}) = 1$; otherwise, $\chi(f_p^{R_i}, f_q^{C_j}) = 0$.

As shown above, we obtain the frame alignment result between two SUs in the reference summary and the candidate summary. Figure 2 shows an example of the result of frame alignment. If a frame in SU_{R_i} (or SU_{C_j}) matches the corresponding frame in SU_{C_j} (or SU_{R_i}), such as $f_5^{R_i}$ and $f_4^{C_j}$, we call it "matched frame"; otherwise, such as $f_3^{R_i}$ and $f_6^{C_j}$, we call it "unmatched frame."

To judge whether SU_{C_j} matches SU_{R_i}, the alignment score of frame alignment is calculated as

$$Align(SU_{R_i}, SU_{C_j}) = s_{m_i n_j}. \tag{4}$$

Considering that SU_{R_i} and SU_{C_j} may partly match, that is, that SU_{C_j} may lose some frames of SU_{R_i} or contain some redundant frames, we calculate the final alignment score as follows:

$$Align(SU_{R_i}, SU_{C_j}) = \frac{1}{\min(m_i, n_j)} s_{m_i n_j}. \tag{5}$$

If the maximal alignment score for SU_{C_j}, according to some SU_{R_i} in all the SUs in the reference summary, is higher than the predefined threshold thr_{SU} ($thr_{SU} = 0.8$ in our experiments), SU_{C_j} is considered to match SU_{R_i}; otherwise, SU_{C_j} is considered as a noise SU. The summary unit matching algorithm is provided in Table 1.

After SU matching, each SU_{C_j} in the candidate summary matches an SU_{R_i} in the reference summary or is considered as a noise SU.

Table 1 Frame alignment-based SU matching between the reference summary and the candidate summary

Algorithm: Summary unit matching

Input: $SU_{C_j} = \{f_1^{C_j}, f_2^{C_j}, \ldots\}_{n_j} \in CS$
$SU_{R_k} = \{f_1^{R_k}, f_2^{R_k}, \ldots\}_{m_k} \in RS, \forall k, k \in \{1, 2, \ldots, N_R\}$

Output: SU_{R_i} or NULL

1. for each $SU_{R_k} = \{f_1^{R_k}, f_2^{R_k}, \ldots\}_{m_k} \in RS$,
   calculate the matching value of each frame pair ($f_p^{R_k}$ and $f_q^{C_j}$),
   match SU_{R_k} and SU_{C_j} using frame alignment:

$$score_{R_k} = Align(SU_{R_k}, SU_{C_j}).$$

2. select $SU_{R_i} \in RS$ with the maximal score:

$$i = \arg \max_{1 \le k \le N_R} (score_{R_k}).$$

3. if $score_{R_i} > thr_{SU}$, return $score_{R_i}$;
   else, return NULL.

Here

RS: the reference summary
CS: the candidate summary
SU_{R_k}: any SU in the reference summary
SU_{C_j}: an SU in the candidate summary
$f_p^{R_k}$: any frame in SU_{R_k}
$f_q^{C_j}$: any frame in SU_{C_j}

4 Similarity-Based Automatic 4C Assessment

The assessment with 4C criteria provides a comprehensive human-like evaluation of video summary quality. It is used to generate the requirement-independent intermediate results as the basis of the further evaluation in our framework. In 4C assessment, we assess the 4C scores of the candidate summary by comparing with the reference summary based on the SU matching result. In the following, we discuss the assessment of coverage, conciseness, coherence, and context, respectively.

4.1 Coverage Assessment

Coverage of the candidate summary represents how much content of the reference summary is covered by the candidate summary.

We define the coverage of the candidate summary as the sum of the coverages of all the SUs in the reference summary:

$$Cov(CS) = \frac{1}{N_R}\sum_{i=1}^{N_R} Cov(SU_{R_i}). \tag{6}$$

For each SU_{R_i} in the reference summary, the coverage of SU_{R_i} is calculated as follows: if none of the SUs in the candidate summary matches SU_{R_i}, $Cov(SU_{R_i})$ is 0; if there only one SU_{C_j} matches SU_{R_i}, $Cov(SU_{R_i})$ is the content of SU_{R_i} covered by SU_{C_j}; if there exist many SUs in the candidate summary that match SU_{R_i}, we choose the SU_{C_j} with the highest alignment score to SU_{R_i} to calculate.

The covered content of SU_{R_i} can be calculated as the sum of the covered content of the frames in SU_{R_i} based on the result of frame alignment in SU matching. As shown in Fig. 2, for a matched frame, such as $f_5^{R_i}$, its covered content can be calculated as the similarity between it and its corresponding frame in alignment. For an unmatched frame, such as $f_3^{R_i}$, its content may be partly covered by the corresponding frames of its nearest matched frame, since the adjacent frames in a video file are usually interrelated in content which calls "temporal redundancy" of video characteristics.

To clearly explain the covered content calculation of SU_{R_i}, we define the concept "related frame." For a matched frame, its related frame is the corresponding frame which matches it. For an unmatched frame, look for the nearest matched frame(s) before or/and after it. If only one matched frame is found, we define the related frame of the found matched frame as the related frame of current unmatched frame; if two matched frames are found, we choose the more similar corresponding frame to current unmatched frame as its related frame. For example, in Fig. 2, $f_5^{R_i}$ matches $f_4^{C_j}$, and the related frame of $f_5^{R_i}$ is $f_4^{C_j}$. $f_3^{R_i}$ does not match any frame in SU_{C_j}, so we look for the nearest matched frame(s) of $f_3^{R_i}$ in SU_{R_i} ($f_2^{R_i}$ and $f_4^{R_i}$) and select the more similar corresponding frame from $f_2^{C_j}$ and $f_3^{C_j}$ as the related frame of $f_3^{R_i}$. In this way, the coverage of SU_{R_i} can be represented as the sum of the similarity between its frames and their related frames. It is calculated as follows:

$$Cov(SU_{R_i}) = \begin{cases} \max_j\left(\frac{1}{m_i}\sum_{p=1}^{m_i} Sim(f_p^{R_i}, RF(f_p^{R_i}))\right) & \text{if } SU_{C_j} \text{ matches } SU_{R_i}, \\ 0 & \text{if no SU matches } SU_{R_i}, \end{cases} \tag{7}$$

where $RF(f_p^{R_i})$ is the related frame of $f_p^{R_i}$ in SU_{C_j}, and $Sim(\cdot,\cdot)$ is the similarity between two frames. In frame similarity measurement, we divide the frames into 4×4 regions with same size and shapes. For each region, 16-bins color histogram on HSV color space is extracted according to MPEG-7 [18]. Each frame is represented by a 256-bins feature vector, and the similarity between two frames is calculated according to Euclidean distance of their feature vectors.

4.2 Conciseness Assessment

Conciseness of the candidate summary represents how much redundant content is contained in the candidate summary.

We define the conciseness of the candidate summary as the sum of the conciseness of all the SUs in the candidate summary:

$$Coc(CS) = \frac{1}{N_C}\sum_{j=1}^{N_C} Coc(SU_{C_j}). \tag{8}$$

For each SU_{C_j} in the candidate summary, the conciseness of SU_{C_j} is calculated as follows: if SU_{C_j} is a noise SU, $Coc(SU_{C_j})$ is 0; if only SU_{C_j} but no other SU in the candidate summary matches a SU_{R_i} in the reference summary, $Coc(SU_{C_j})$ is the useful content of SU_{C_j} which is also contained by SU_{R_i}; if there exist many SUs in the candidate summary which match the same SU_{R_i} in the reference summary, we select the SU_{C_j} with the highest alignment score to SU_{R_i} to calculate and consider that the concisenesses of the other unselected SUs are 0.

Similar to coverage assessment, conciseness of SU_{C_j} is calculated as the sum of the contained useful content in frames of SU_{C_j}. Based on the result of frame alignment in SU matching, the useful content contained in a frame of SU_{C_j} is calculated as the similarity between it and its related frame in SU_{R_i}, and the conciseness of SU_{C_j} is calculated as follows:

$$Coc(SU_{C_j}) = \begin{cases} \frac{1}{n_j}\sum_{q=1}^{n_j} Sim(f_q^{C_j}, RF(f_q^{C_j})) \\ \quad \text{if } SU_{C_j} \text{ is the selected SU matching } SU_{R_i}, \\ 0 \quad \text{if } SU_{C_j} \text{ is a noise or unselected SU,} \end{cases} \tag{9}$$

where $RF(f_q^{C_j})$ is the related frame of $f_q^{C_j}$ in SU_{R_i}, and $Sim(\cdot,\cdot)$ is defined as in (7).

4.3 Coherence Assessment

Coherence of the candidate summary represents how coherent is the candidate summary in representation.

We consider the coherence of the candidate summary in two aspects: inter SU coherence and inner SU coherence. Inter SU coherence means the coherence between SUs, and inner SU coherence means the coherence within each SU. The coherence of the candidate summary is defined as follows:

$$Coh(CS) = \omega_1 \cdot Coh_{\text{inner}}(CS) + \omega_2 \cdot Coh_{\text{inter}}(CS), \tag{10}$$

where ω_1 and ω_2 are positive weight coefficients, and $\omega_1 = \omega_2 = 0.5$ in our experiments.

We first assess the inter SU coherence by comparing the mean values of the distances between two adjacent SUs in the reference summary and the candidate

summary. The distance between two adjacent SUs in a video summary S is calculated as the distance between the last frame of the former SU and the first frame of the latter SU:

$$Dis_{\text{inter}}(SU_{S_k}, SU_{S_{k+1}}) = \mathcal{D}\big(f_{n_k}^{S_k}, f_1^{S_{k+1}}\big), \tag{11}$$

where SU_{S_k} and $SU_{S_{k+1}}$ are two adjacent SUs in video summary S; $f_{n_k}^{S_k}$ and $f_1^{S_{k+1}}$ are the last frame of SU_{S_k} and the first frame of $SU_{S_{k+1}}$, respectively; and $\mathcal{D}(\cdot,\cdot)$ denotes the distance between two frames, which is calculated by Euclidean distance of their local HSV color histogram feature vectors.

The inter SU coherence is calculated as follows:

$$\begin{aligned} Coh_{\text{inter}}(CS) \;=\; 1 - \max\Bigg(0,\, &\frac{1}{N_C - 1}\sum_{j=1}^{N_C-1} Dis_{\text{inter}}(SU_{C_j}, SU_{C_{j+1}}) \\ &- \frac{1}{N_R - 1}\sum_{i=1}^{N_R-1} Dis_{\text{inter}}(SU_{R_i}, SU_{R_{i+1}})\Bigg). \end{aligned} \tag{12}$$

Next, we define the inner SU coherence of the candidate summary as the sum of the inner coherences of all SUs:

$$Coh_{\text{inner}}(CS) = \frac{1}{N_C}\sum_{j=1}^{N_C} Coh_{\text{inner}}(SU_{C_j}). \tag{13}$$

The inner coherence of SU_{C_j} is calculated by comparing to its matching SU_{R_i} in the reference summary. To calculate the inner coherence of each SU, we define the "average distance" between two frames f_p and f_q as follows:

$$\widetilde{\mathcal{D}}(f_p, f_q) = \begin{cases} \frac{1}{q-p}\sum_{k=p}^{q-1} \mathcal{D}(f_k, f_{k+1}), & p < q, \\ 0, & p \geq q, \end{cases} \tag{14}$$

where $\widetilde{\mathcal{D}}(,)$ denotes the average distance between two frames.

Then, we evaluate the inner SU coherence by comparing the distance between each frame and its successive frame with the average distance between their related frames:

$$\begin{aligned} Coh_{\text{inner}}(SU_{C_j}) \;=\; 1 - \frac{1}{n_j - 1}\sum_{k=1}^{n_j-1} \max\Big(0, \mathcal{D}\big(f_k^{C_j}, f_{k+1}^{C_j}\big) \\ - \widetilde{\mathcal{D}}\big(RF\big(f_k^{C_j}\big), RF\big(f_{k+1}^{C_j}\big)\big)\Big), \end{aligned} \tag{15}$$

where $RF(f_k^{C_j})$ is the related frame of $f_k^{C_j}$ in SU_{R_i}.

Note here that a noise SU in the candidate summary does not have a matching SU in the reference summary and its frames do not have their related frames. Hence,

we replace the average distance between the related frames in (15) with the mean value of the distances between each frame and its successive frame in all the SUs of the reference summary. The inner coherence of a noise SU is calculated as

$$Coh_{\text{inner}}(SU_{C_j}) = 1 - \frac{1}{n_j - 1}\sum_{k=1}^{n_j-1}\max\left(0, \mathcal{D}\left(f_k^{C_j}, f_{k+1}^{C_j}\right) - \frac{1}{N_R}\sum_{i=1}^{N_R}\widetilde{\mathcal{D}}\left(f_1^{R_i}, f_{m_i}^{R_i}\right)\right), \tag{16}$$

where $f_1^{R_i}$ and $f_{m_i}^{R_i}$ are the first and the last frames of SU_{R_i}.

4.4 Context Assessment

Context of the candidate summary represents how ordered the SUs of the candidate summary are.

Since the noise SUs and the missing SUs do not influence the order of the other SUs in the candidate summary, we ignore them in context assessment. For the repeated SUs in the candidate summary, that is, when more than one SU matches the same SU in the reference summary, we retain one of the SUs in the candidate summary each time and compute the mean value of its contexts in all situations. So the context of the candidate summary is defined as follows:

$$Cot(CS) = \frac{1}{N_S}\sum_{k=1}^{N_s} Cot_k(CS), \tag{17}$$

$$N_S = \prod_{i=1}^{N_R}\max(1, N_i), \tag{18}$$

where N_i is the number of SUs in the candidates summary that match SU_{R_i} in SU matching, and N_S is the number of all possible situations.

To calculate the context of the candidate summary, we define the order of SUs. If SU_{S_i} and SU_{S_j} are two SUs in a video summary S, then the order of SU_{S_i} and SU_{S_j} is

$$O_S(SU_{S_i}, SU_{S_j}) = \begin{cases} 1 & \text{if } SU_{S_i} \text{ appears in front of } SU_{S_j} \text{ in } S, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

For SU_{C_j} and SU_{C_q} in the candidate summary, we define the "inversion" as follows:

$$Inv(SU_{C_j}, SU_{C_q}) = \begin{cases} 1, & O_{CS}(SU_{C_j}, SU_{C_q}) \neq O_{RS}(SU_{R_i}, SU_{R_p}), \\ 0, & O_{CS}(SU_{C_j}, SU_{C_q}) = O_{RS}(SU_{R_i}, SU_{R_p}), \end{cases} \tag{20}$$

where SU_{C_j} and SU_{C_q} match SU_{R_i} and SU_{R_p} in SU matching, respectively.

We define the context of the candidate summary as follows:

$$Cot_k(CS) = 1 - \frac{\sum_{j \neq q} \mathcal{E}(SU_{C_j}, SU_{C_q}) \cdot Inv(SU_{C_j}, SU_{C_q})}{\sum_{j \neq q} \mathcal{E}(SU_{C_j}, SU_{C_q})}, \tag{21}$$

where $\mathcal{E}(SU_{C_j}, SU_{C_q})$ is the effect of SU_{C_j} to the understanding of SU_{C_q}.

In this chapter, we assume that the viewer will not trace back and only consider the effect of the prior SUs to the understanding of the following SUs. We consider the effect of SU_{C_j} to the understanding of SU_{C_q} to be determined by the distance between their matched SUs in the reference summary and calculated as follows:

$$\mathcal{E}(SU_{C_j}, SU_{C_q}) = \begin{cases} F(|p-i|), & O_{RS}(SU_{R_i}, SU_{R_p}) = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{22}$$

where F is a decreasing function, e.g., $F(x) = 1/x$.

According to the above methods, we can obtain the scores of the candidate summary on 4C criteria. But these scores may not exactly match the manual evaluation results though they are highly related to user perception. So, we fit the scores to generate the final 4C automatic assessment results.

We utilize linear regression to fit the score on each criterion:

$$\begin{pmatrix} s_{Cov} \\ s_{Coc} \\ s_{Coh} \\ s_{Cot} \end{pmatrix} = \begin{pmatrix} \alpha_{Cov} \\ \alpha_{Coc} \\ \alpha_{Coh} \\ \alpha_{Cot} \end{pmatrix} + \mathrm{diag}(\beta_{Cov}, \beta_{Coc}, \beta_{Coh}, \beta_{Cot}) \begin{pmatrix} Cov(CS) \\ Coc(CS) \\ Coh(CS) \\ Cot(CS) \end{pmatrix}, \tag{23}$$

where s_{Cov}, s_{Coc}, s_{Coh}, s_{Cot} are the final 4C assessment scores, and α_{Cov}, α_{Coc}, α_{Coh}, α_{Cot}, β_{Cov}, β_{Coc}, β_{Coh}, β_{Cot} are the weight coefficients.

These weight coefficients can be calculated by the least squares method:

$$(\alpha_{\#}, \beta_{\#}) = \arg\min \sum \left(s'_{\#} - \alpha_{\#} - \beta_{\#} \cdot \#(CS)\right)^2, \tag{24}$$

where # is a criterion in 4C criteria, including *Cov*, *Coc*, *Coh*, and *Cot*; $s'_{\#}$ is the manual evaluation result on the criterion; #(*CS*) is the automatic assessment score on the criterion before fitting.

5 User Interaction Based Individual Evaluation

Though 4C criteria can provide comprehensive description of video summary quality, the viewpoint and perspective of video summary quality are usually application-dependent [29]. This means that the users may require individual evaluation results with various criteria in different applications. For example, the rushes summarization task in TRECVID 2007 requires evaluating video summary quality in "INclusion of ground truth" (IN), "EAse of understanding" (EA), and "lack of redundancy"

(RE) [22]. Designing the automatic assessment methods for each required evaluation criterion as above will lead to high labor cost of experts, and the existing automatic assessment methods cannot be well reused when new criteria are required. In our framework, we propose an effective approach to transform the automatic 4C assessment results to the evaluation results satisfying user requirements. For any required criteria, the approach can build the transformation model between 4C criteria and the required criteria with some user interaction, and automatically transform the 4C assessment scores to the required evaluation results.

5.1 User Interaction Based Requirement Gathering

In the procedure of building the transformation model, we first gather the user requirements of video summary quality by means of user interaction. The training data with limited size is generated, and the automatic 4C assessment scores and the manual evaluation results with the required criteria on the training data are used to build the transformation model.

To gather the user requirements, we ask the users to evaluate the training data with their criteria. In the user interaction, each user watches a reference summary for three times to make the video content familiar and evaluates the corresponding candidate summaries in a random order. To eliminate the influence of evaluation order, the first evaluated candidate summary for each video file will be evaluated again. When evaluating a candidate summary, the users are allowed to watch the reference summary again but forbidden any operation in the candidate summary playing. Figure 3 shows the interface used in user interaction. The reference summary and the candidate summary are displayed in the top of the interface. When evaluating a candidate summary, the user can choose to watch the reference summary first (press the left button with the text "play *RS*") or directly watch the candidate summary (press the right button with the text "play *CS*"). If the user chooses to watch the reference summary first, the candidate summary will be automatically played following the reference summary. After watching the candidate summary, the users are asked to input the quality scores according to his/her required criteria, from 1 to M representing the quality from the worst to the best in the corresponding criterion. The textboxes for inputting evaluation results are in the bottom of the interface. Since the number of the required criteria may be variable in different evaluations, the required criteria are shown available (in white color) and marked with the corresponding criteria labels (such as "IN", "EA", "RE").

5.2 Transformation of 4C Assessment Scores

When obtaining the user interaction results, we build the transformation model for adapting the 4C assessment scores to the required individual evaluation results. We

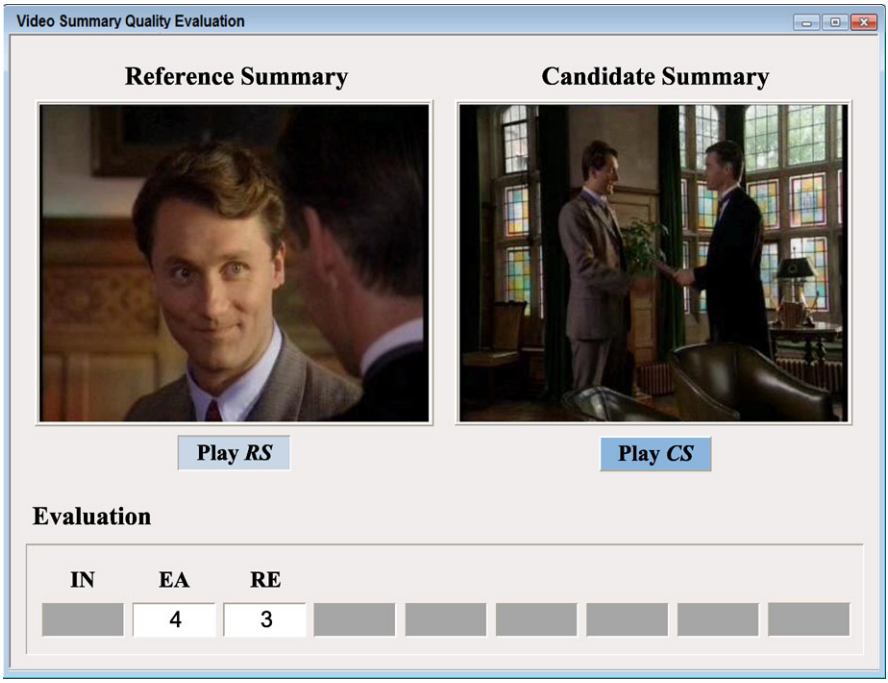

Fig. 3 User interface for manual evaluation of video summary quality

use the weighted-sum model in this chapter, and more complex transformation models are left for the future work.

We represent the scores of the required evaluation criteria as

$$\mathbf{G} = (g_1, g_2, \ldots, g_N)^T, \tag{25}$$

where N is the number of aspects in the required evaluation criteria.

Then we calculate the elements of $\mathbf{G}$ by the weighted sums of the 4C assessment scores. Each g_j in $\mathbf{G}$ can be represented as

$$g_j = 1 + (M-1) \cdot (\lambda_{j0} + \lambda_{j1} s_{Cov} + \lambda_{j2} s_{Coc} + \lambda_{j3} s_{Coh} + \lambda_{j4} s_{Cot}), \tag{26}$$

where $\lambda_{j0}, \lambda_{j1}, \ldots, \lambda_{j4}$ are the weight coefficients for evaluation result transformation. We constrain the value of λ_{j0} in the range of [–1,1] and the values of the other coefficients in the range of [0, 1], and $\lambda_{j1} + \lambda_{j2} + \lambda_{j3} + \lambda_{j4} = 1$. The constants 1 and $(M-1)$ are used to ensure g_j in the range of $[1, M]$.

To simplify representation, let $\mathbf{X} = (x_0, x_1, x_2, x_3, x_4)^T$ and $\mathbf{Y} = (y_1, y_2, \ldots, y_N)^T$, where $x_0 = 1$, x_1 to x_4 denote s_{Cov}, s_{Coc}, s_{Coh}, s_{Cot} in that order, and $y_j = (g_j - 1)/(M-1)$. Since the 4C assessment scores are in the range of [0,1] and the evaluation scores with the required criteria are from 1 to M, each element

Table 2 Transformation model from the 4C assessment scores to the individual evaluation results

Algorithm: Transforming 4C assessment scores to the required evaluation results

Input: 4C assessment scores on the training data
manual evaluation results with the required criteria on the training data

Output: weight coefficient matrix Λ

1. Generate the observation matrix $\widehat{\mathbf{X}}$ from 4C assessment scores

$$\widehat{\mathbf{X}} = \begin{bmatrix} x_{01} & x_{02} & \cdots & x_{0r} \\ x_{11} & x_{12} & \cdots & x_{1r} \\ \vdots & \vdots & & \vdots \\ x_{41} & x_{42} & \cdots & x_{4r} \end{bmatrix}.$$

2. Generate the observation matrix $\widehat{\mathbf{Y}}$ from the manual evaluation results with the required criteria

$$\widehat{\mathbf{Y}} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1r} \\ y_{21} & y_{22} & \cdots & y_{2r} \\ \vdots & \vdots & & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{Nr} \end{bmatrix}.$$

3. Calculate the weight coefficient matrix Λ by multivariate linear regression

$$\lambda_{\mathbf{j}} = (\lambda_{j0}, \lambda_{j1}, \ldots, \lambda_{j4}) = \arg\min \sum_{k=1}^{r} (y_{jk} - \lambda_{j0}x_{0k} - \lambda_{j1}x_{1k} - \cdots - \lambda_{j4}x_{4k})^2,$$

$$\Lambda = (\lambda_{\mathbf{1}}; \lambda_{\mathbf{2}}; \ldots; \lambda_{\mathbf{N}}).$$

in $\mathbf{X}$ and $\mathbf{Y}$ is in the range of [0, 1]. Then, (26) can be represented as

$$y_j = \lambda_{\mathbf{j}}\mathbf{X} = \lambda_{j0}x_0 + \lambda_{j1}x_1 + \cdots + \lambda_{j4}x_4. \tag{27}$$

Considering each variable y_j in $\mathbf{Y}$ separately, we calculate the transformation between $\mathbf{X}$ and $\mathbf{Y}$ by multivariate linear regression. Assuming that there are r independent observations of y_j, the best weight coefficient vector λ_j for y_j can calculated by the least squares method:

$$(\lambda_{j0}, \lambda_{j1}, \ldots, \lambda_{j4}) = \arg\min \sum_{k=1}^{r} (y_{jk} - \lambda_{j0}x_{0k} - \lambda_{j1}x_{1k} - \cdots - \lambda_{j4}x_{4k})^2, \tag{28}$$

where $(x_{0k}, \ldots, x_{4k}, y_{jk})$ denotes the kth observation of y_j.

Table 2 shows the procedure of calculating the weight coefficient matrix for transforming the 4C assessment scores to the required evaluation results.

With multivariate linear regression, an $N \times 5$ weight coefficient matrix Λ will be generated. In matrix Λ, the coefficients in each row represent the weights of

the 4C assessment scores in calculating the result of the corresponding required criterion, and the sum of the coefficients in each column represents the weight of the corresponding 4C criterion in the required evaluation criteria. The evaluation results according to user requirements can be transformed from 4C assessment results as follows:

$$\mathbf{G} = 1 + (M - 1) \cdot \Lambda \mathbf{X}. \tag{29}$$

5.3 Incremental User Interaction

The complexity of correlations between the 4C assessment scores and the different required criteria are usually different. For example, the IN and RE criteria used in TRECVID 2007 have direct correlation to the coverage and conciseness criteria, respectively, but the EA criterion has more complex correlation to 4C criteria. For these required criteria, the sizes of training data to build their transformation models may be different.

In order to reduce the labor cost in user interaction, we carry out the user interaction in an incremental way. Initially, a subset of the training data is selected by random sampling in the 4C assessment score space. After the evaluators evaluate the subset and the corresponding weight coefficient matrix is generated, we calculate the mean absolute error (MAE) in each criterion on the subset. Since the subset is selected by random sampling, we consider that the weight coefficients can well transform the 4C assessment scores to the required evaluation results if the MAE in some criterion is smaller than a predefined threshold,\ and stop the evaluators to further evaluate in this criterion by setting the corresponding textbox to unavailable, such as the IN criterion in Fig. 3. For the remaining criteria, we incrementally provide more candidate summaries by randomly sampling in the training data till the MAEs in all criteria are smaller than the predefined threshold or all the candidate summaries in the training data are evaluated.

6 Experiments

We validate the performance of the proposed full-reference evaluation techniques for video summary on the standard dataset from TRECVID 2007 rushes summarization task. There are three reasons to select this dataset for our experiment. First, as a global competition in video summarization, TRECVID rushes summarization task provides an accepted dataset and the corresponding video summaries generated by different participants for each original video, which can be used to generate the reference summaries and the candidate summaries. Second, the rushes videos are the unedited raw footages with several repeats of each shot, so the summaries generated from rushes usually have more problems in redundancy and context than the summaries generated from other videos. It can more efficiently validate the performance

of 4C assessment algorithm. Third, TRECVID provides the criteria to evaluate the generated video summaries that can be used to validate the user interaction-based individual evaluation method.

While building the dataset in our experiments, we select ten rushes from 42 files that have different source files, durations, retake times, and movie tones. The selected rushes are: MRS025913, MRS042543, MRS042548, MRS043400, MRS044500, MRS048779, MRS145918, MRS157445, MRS157475, MS210470. Each video file used in our experiments is generated from one selected rush, and it includes one typical shot with multiple retakes. The reference summary of each video file is generated by manually assembling the extracted frames. We also select ten participants from total twenty-four participants, whose provided summaries include the corresponding parts of our selected video shot files and have different performances in the competition. The ten selected participants are: attlabs, cityu, cmu, cost292, hkpu, kddietal, ntu, thu-icrc, ucal, umadrid. Each participant provides one candidate summary for each video file in the experiments, so totally there are one hundred candidate summaries. We randomly select fifty candidate summaries to build the training data and treat the rest fifty candidate summaries as the test data. To provide the manual evaluation results, we invite ten volunteers as the evaluators in our user studies. They are in age of 20 to 40, including undergraduate and graduate students, officers, and company employees. To our knowledge, they have no idea about our work before the user studies. To eliminate the personal evaluation preference, the mean value of the evaluation results from all evaluators to the same candidate summary in each criterion is treated as the final manual evaluation result of the candidate summary in this criterion.

In this section, the first experiment provides the validation of the 4C assessment algorithm, the second experiment presents the feasibility of the incremental user interaction, and the third experiment shows how to effectively transform the 4C assessment scores to the evaluation results with the required criteria.

6.1 Validation of 4C Assessment Algorithm

We first demonstrate the 4C assessment algorithm on shot 103 in rushes file MRS044500, which has been chosen as the demo video in the TRECVID 2007 for rushes summarization task.

The reference summary is generated manually, and eight candidate summaries are described in Table 3. We decompose the reference summary and the candidate summaries to a set of SUs as shown in Fig. 4 and calculate the scores of candidate summaries' quality by 4C assessment.

Table 4 shows the 4C assessment results of the candidate summaries. Candidate summary 1 obtains full scores in all four criteria because it is totally same with the reference summary. Candidate summaries 2 to 5 are four artificially generated summaries with the obvious problems in coverage, conciseness, coherence, and context, respectively. Candidate summary 2 misses the last two SUs of reference summary,

Fig. 4 Video summaries for the rushes file of shot 103 in MRS044500. Here, the SUs in the reference summary (*RS*) are represented with $SU_1, \ldots, SU_8$. The SUs in the candidate summaries are represented according the SUs in the references summary: SU_i denotes a same SU to the SU_i in *RS*; $RDSU_i$ denotes a reduced SU of the SU_i in *RS*; $RTSU_i$ denotes a retake of the SU_i in *RS*; $NRSU_i$ denotes a near SU of the SU_i in *RS*, which can be a retake, a reduced one, or any other similar one; SU_{noise} denotes a noise SU

Table 3 Different candidate video summaries for the rushes file of shot 103 in MRS044500

CS No.	Description of the candidate summary
CS 1	same with the reference summar
CS 2	remove the last two SUs from the reference summary
CS 3	add two noise SUs in the head and end of the reference summary
CS 4	drop the first 20% and the last 20% frames of each SU in the reference summary
CS 5	invert the orders of the SUs in the reference summary
CS 6	a retake of the reference summary
CS 7	baseline summary (select 1 second in each 25 seconds of the original video)
CS 8	a summary from one participant in TRECVID 2007

Table 4 4C assessment scores on shot 103 in MRS044500

CS No.	s_{Cov}	s_{Coc}	s_{Coh}	s_{Cot}
CS 1	1.000	1.000	1.000	1.000
CS 2	0.750	1.000	1.000	1.000
CS 3	1.000	0.800	1.000	1.000
CS 4	0.954	1.000	0.889	1.000
CS 5	1.000	1.000	1.000	0.652
CS 6	0.904	0.906	0.903	1.000
CS 7	0.690	0.421	0.827	0.541
CS 8	0.713	0.408	0.765	0.580

so the coverage is poor. Similarly, candidate summary 3 has two noise SUs in the head and end, so the conciseness is poor. Candidate summary 4 is generated by dropping 20% frames at the beginning of each SU and 20% at the end of each SU; therefore, it leads to incoherence. In candidate summary 5, the SU sequence has the wrong order, so the score of context is low. Candidate summary 6 is a retake of the reference summary, so it has good performance in all four criteria. Candidate summary 7 is one baseline summary of TRECVID 2007, and candidate summary 8 is the summary from one participant. These two candidate summaries are generated by automatic multimedia content analysis algorithms. Obviously, their performances are not as good as the artificially generated summaries, and the problems of quality are more complicated.

To further validate the effectiveness of the 4C assessment algorithm, we carry out a user study on the whole dataset. We explain the 4C criteria to the evaluators for five minutes before the manual evaluation. Then, each evaluator is asked to evaluate all the candidate summaries according to 4C criteria. The evaluation results in the value range from 0 to 1 with ten steps, and higher score means better performance.

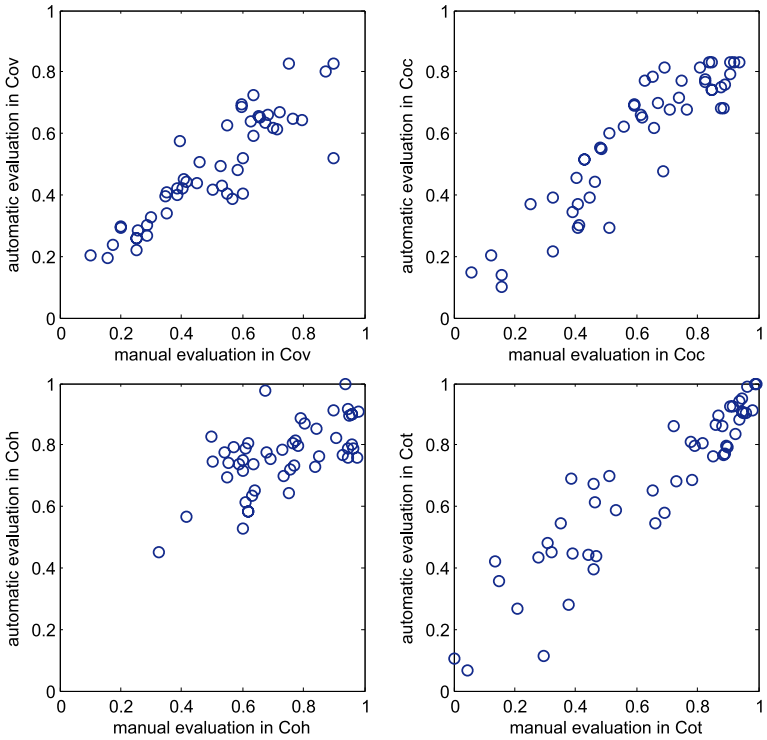

Fig. 5 Comparison of manual evaluation results and automatic assessment results according to 4C criteria on the test data

Table 5 Performance of automatic 4C assessment on the test data

	s_{Cov}	s_{Coc}	s_{Coh}	s_{Cot}
MAE	0.064	0.083	0.107	0.085
CC	0.897	0.915	0.626	0.926

Figure 5 shows a comparison between the manual evaluation results and the automatic assessment results according to 4C criteria on the test data. Table 5 shows the MAE of the automatic 4C assessment results and their correlation coefficients (CC) to the manual evaluation results. It shows high correlation in coverage, conciseness, and context between the manual evaluation results and automatic assessment results. The results in coherence show weaker correlation, since the evaluators usually hardly give very accurate judgments to the intensity and frequency of incoherence in evaluation.

6.2 Validation of Incremental User Interaction

In this subsection, we present the feasibility of incremental user interaction. We select the human-like criteria in TRECVID 2007 rushes summarization task as the required criteria, including IN, EA, and RE [22].

Since the provided scores of IN, EA, and RE in TRECVID 2007 are given to the summaries of the total rushes but not the typical shots, we ask the ten evaluators to evaluate the candidate summaries with the proposed user interaction approach in Sect. 5. The scores used in evaluation are in five levels, i.e., from 1 to 5.

Using the transformation model generation algorithm in Table 2, we generate the transformation model from 4C assessment scores to the required evaluation results on the training data. We initially build a subset with 40% size of the training data (20 candidate summaries) and incrementally add 10 candidate summaries every time. Table 6 shows the generated weight coefficient matrix in each step. We can find that the criteria directly correlated to 4C criteria, such as IN and RE, can reach the stable weight coefficients rapidly, and the criteria with more complex correlation to the 4C assessment scores, such as EA, require more training data to adjust the corresponding transformation models.

In our experiments, we use 0.1 as the threshold to measure the mean absolute error on the training data. Hence, only the manual evaluation in EA is carried out on the whole training data, and the manual evaluation in IN and RE is stopped after the evaluation on the initial subset with 40% size. It shows that the proposed incremental approach can reduce the labor cost in user interaction.

Table 6 Weight coefficient matrix in incremental user interaction

λ_{ji}	x_0	x_1	x_2	x_3	x_4	MAE	
$k = 40\%$	y_1	**0.073**	**0.985**	**0.007**	**0.009**	**0.018**	**0.083**
	y_2	0.058	0.381	0.042	0.028	0.524	0.137
	y_3	**0.025**	**0.009**	**0.967**	**0.023**	**0.004**	**0.091**
$k = 60\%$	y_1	0.135	0.962	0.016	0.007	0.002	0.084
	y_2	0.039	0.363	0.071	0.032	0.513	0.162
	y_3	0.047	0.006	0.953	0.016	0.005	0.083
$k = 80\%$	y_1	0.064	0.971	0.021	0.008	0.009	0.086
	y_2	0.048	0.337	0.076	0.061	0.540	0.131
	y_3	0.053	0.012	0.957	0.014	0.003	0.095
$k = 100\%$	y_1	0.079	0.976	0.013	0.003	0.006	0.085
	y_2	**0.061**	**0.344**	**0.051**	**0.019**	**0.541**	**0.107**
	y_3	0.052	0.018	0.968	0.009	0.007	0.092

6.3 Validation of Evaluation Result Transformation

In this subsection, we will show how to transform the 4C assessment scores to satisfy various evaluation requirements.

Based on incremental user interaction and the transformation generation algorithm in Table 2, we can calculate the weight coefficient matrix to transform the 4C assessment scores. Table 7 gives the weight coefficient values for IN, EA, and RE. From Table 7 we can find that coverage and conciseness dominate the IN and RE scores, respectively, while coverage and context dominate the EA score together. It is consonant with the definition of IN, EA, RE in TRECVID 2007 [22].

We generate the automatic evaluation results in IN, EA, RE by weighted sum of the 4C assessment scores. Figure 6 shows the comparison between the automatic evaluation results and manual evaluation results in IN, EA, RE on the test data. Table 8 shows the MAEs of the automatic evaluation results and their correlation coefficients to the manual evaluation results. It is obvious that the proposed automatic evaluation techniques can fit manual evaluation very well.

We further calculate the sum of coefficients in each column as the total weight of the corresponding criterion in 4C criteria (Table 2). It can be used to assess the criteria used in video summary quality evaluation. It shows that coverage and conciseness are fully considered in the evaluation criteria including IN, EA, RE, but the coherence is ignored.

Table 7 Weight coefficient matrix for individual evaluation results generation

λ_{ji}	$x_0 = 1$	s_{Cov}	s_{Coc}	s_{Coh}	s_{Cot}
g_{IN}	0.073	0.985	0.007	0.009	0.018
g_{EA}	0.061	0.344	0.051	0.019	0.541
g_{RE}	0.025	0.009	0.967	0.023	0.004
total weight		1.338	1.025	0.051	0.563

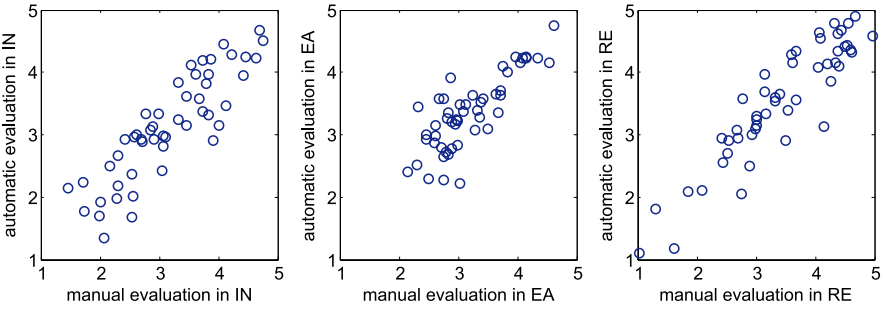

Fig. 6 Comparison of manual evaluation results and automatic evaluation results in IN, EA, RE on the test data

Table 8 Performance of automatic IN, EA, RE evaluation on the test data

	g_{IN}	g_{EA}	g_{RE}
MAE	0.349	0.467	0.328
CC	0.875	0.813	0.906

7 Conclusions

This chapter presents a novel framework to evaluate the quality of the video summary according to various user requirements. The framework replies on three underlying algorithms that are well adapted to the characteristics of video summary evaluation: frame alignment-based summary unit matching, similarity-based automatic 4C assessment, and incremental user interaction-based individual evaluation. Together, they provide a complete evaluation framework that well satisfies the user requirements in video summary quality evaluation. We have illustrated the performance of proposed techniques on the standard dataset of rushes summarization in TRECVID 2007.

Further work will be explored from two aspects. First, we intend to seek the quality evaluation method without the requirement of a perfect reference summary, i.e., nonreference or reduced-reference evaluation for video summary. Second, current transformation model is based on linear combination of 4C assessment scores. We will consider the possibility of other models and compare the evaluation performance with the weighted sum model.

Acknowledgements We would like to thank the volunteers for their work in manual evaluation. This work is supported by the National Natural Science Foundation of China (60721002, 60975043) and Hong Kong General Research Fund PolyU 5204/09E.

References

1. Agnihotri, L., Dimitrova, N., Kender, J.R.: Design and evaluation of a music video summarization system. In: IEEE International Conference on Multimedia and Expo, Taipei (2004)
2. Das, D., Martins, A.F.T.: A survey on automatic text summarization. CMU Course (2007). http://www.cs.cmu.edu/dipanjan/pubs/summarization. Accessed 20 February 2010
3. Du, Y., Shao, L.: Video shots retrieval using local invariant features. In: International Workshop on Interactive Multimedia for Consumer Electronics, Beijing (2009)
4. Dumont, E., Mérialdo, B.: Rushes video summarization and evaluation. Multimed. Tools Appl. (2009). doi:10.1007/s1104200903749
5. Gong, Y., Liu, X.: Summarizing video by minimizing visual content redundancies. In: IEEE International Conference on Multimedia and Expo, Tokyo (2001)
6. He, L., Sanocki, E., Gupta, A., Grudin, J.: Auto-summarization of audio-video presentations. In: ACM International Conference on Multimedia, Florida (1999)
7. Hua, X.S., Liu, W., Zhang, H.J.: An automatic performance evaluation protocol for video text detection algorithms. IEEE Trans. Circuits Syst. Video Technol. **14**(4), 498–507 (2004)
8. Huang, M., Mahajan, A.B., DeMenthon, D.F.: Automatic performance evaluation for video summarization. Technique Report of Maryland University, Maryland (2004)

9. Knoche, H., De Meer, H.G., Kirsh, D.: Utility curves: mean opinion scores considered biased. In: International Workshop Quality of Service, London (1999)
10. Kusuma, T.M., Zepernick, H.J.: A reduced-reference perceptual quality metric for in-service image quality assessment. In: Mobile Future and Symposium Trends in Communications, Bratislava (2003)
11. Law-To, J., Chen, L., Joly, A., Laptev, I., Buisson, O., Gouet-Brunet, V., Boujemaa, N., Stentiford, F.: Video copy detection: a comparative study. In: International Conference on Image and Video Retrieval, Amsterdam (2007)
12. Li, Y., Zhang, T., Tretter, D.: An overview of video abstraction techniques. HP Technique Report (2001)
13. Lienhart, R., Effelsberg, W., Jain, R.: VisualGREP: a systematic method to compare and retrieve video sequences. Multimed. Tools Appl. **10**(1), 47–72 (2000)
14. Liu, T., Zhang, H.J., Qi, F.: A novel video key-frame extraction algorithm based on perceived motion energy model. IEEE Trans. Circuits Syst. Video Technol. **13**(10), 1006–1013 (2003)
15. Liu, Y., Zhang, Y., Sun, M., Li, W.: Full-reference quality diagnosis for video summary. In: IEEE International Conference Multimedia and Expo, Hannover (2008)
16. Lowe, D.G.: Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision, Kerkyra (1999)
17. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: ACM International Conference on Multimedia, Juan les Pins (2002)
18. Manjunath, B.S., Ohm, J.R., Vasudevan, V.V., Yamada, A.: Color and texture descriptors. IEEE Trans. Circuits Syst. Video Technol. **11**(6), 703–715 (2001)
19. Naphide, H.R., Huang, T.S.: A probabilistic framework for semantic video indexing, filtering, and retrieval. IEEE Trans. Multimed. **3**(1), 141–151 (2001)
20. Naturel, X., Gros, P.: A fast shot matching strategy for detecting duplicate sequences in a television stream. In: International Workshop on Computer Vision Meets Databases, Maryland (2005)
21. Needleman, S., Wunsch, C.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**(3), 443–453 (1970)
22. Over, P., Smeaton, A.F., Kelly, P.: The TRECVid 2007 BBC rushes summarization evaluation pilot. In: ACM International Workshop on TRECVid Video Summarization, Maryland (2007)
23. Over, P., Smeaton, A.F., Awad, G.: The TRECVid 2008 BBC rushes summarization evaluation. In: ACM International the Workshop on TRECVid Video Summarization, Vancouver (2008)
24. Ren, T., Liu, Y., Wu, G.: Full-reference quality assessment for video summary. In: International Workshop on Video Mining, Pisa (2008)
25. Sheikh, H.R., Wang, Z., Cormack, L., Bovik, A.C.: Blind quality assessment for JPEG2000 compressed images. In: International Conference on Signals, Systems and Computers, California (2002)
26. Shen, H.T., Ooi, B.C., Zhou, X.: Towards effective indexing for very large video sequence database. In: ACM International Conference on Management of Data, Maryland (2005)
27. Silva, G.C., Yamasaki, T., Aizawa, K.: Evaluation of video summarization for a large number of cameras in ubiquitous home. In: ACM International Conference on Multimedia, Singapore (2005)
28. Taskiran, C.M.: Evaluation of automatic video summarization systems. In: International Conference on Multimedia Content Analysis, Management, and Retrieval, California (2006)
29. Truong, B.T., Venkatesh, S.: Video abstraction: a systematic review and classification. ACM Trans. Multimed. Comput. Commun. Appl. **3**(1), 1–37 (2007)
30. Uchihashi, S., Foote, J., Girgensohn, A., Boreczky, J.: Video Manga: generating semantically meaningful video summaries. In: ACM International Conference on Multimedia, Florida (1999)

31. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
32. Yahiaoui, I., Merialdo, B., Huet, B.: Comparison of multiepisode video summarization algorithms. EURASIP J. Appl. Signal Process. **2003**(1), 48–55 (2003)
33. Zhang, H.J., Kankanhalli, A., Smoliar, S.W.: Automatic partitioning of full-motion video. In: Jeffay, K., Zhang, H.J. (eds.) Readings in Multimedia Computing and Networking, 1st edn. Morgan Kaufmann, San Francisco (2001)