2008 IEEE International Conference on Data Mining Workshops

Full-reference Quality Assessment for Video Summary

Tongwei Ren[1,2], Yan Liu[2], Gangshan Wu[1]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, 210093, P.R.China
[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
rentw@graphics.nju.edu.cn, csyliu@comp.polyu.nju.hk, gswu@graphics.nju.edu.cn

Abstract

As video summarization techniques have attracted more and more attention for efficient multimedia data management, quality assessment of video summary is required. To address the lack of automatic evaluation techniques, this paper proposes a novel framework including several new algorithms to assess the quality of the video summary against a given reference. First, we partition the reference video summary and the candidate video summary into the sequences of Summary Unit (SU). Then, we utilize alignment based algorithm to match the SUs in the candidate summary with the SUs in the corresponding reference summary. Third, we propose a novel similarity based 4C-assessment algorithm to evaluate the candidate video summary from the perspective of coverage, conciseness, coherence, and context, respectively. Finally, the individual assessment results are integrated according to user's requirement by a learning based weight adaptation method. The proposed framework and techniques are experimented on a standard dataset of TRECVID 2007 and show the good performance in automatic video summary assessment.

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 management [8]. 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. A good video summary epitomizes the essentials of the original video in the form of storyboard (a collection of still images), or video skim (a much shorter vide clip) [7]. An informative and concise video summary enables efficient access to the voluminous, redundant and unstructured video collections [1].

Although video summarization has received more and more attention, a consistent evaluation framework for video summarization is still unavailable [14]. Currently, the quality of the video summary is mainly assessed by human individuals [2, 10, 13], which is seriously influenced by human factors. Moreover, this kind of subjective evaluation has high labor cost and time cost [10]. The missing of the objective assessment 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 [14].

Due to the limitation of subjective evaluation for video summary, objective assessment techniques providing the human like evaluation are highly demanded [4]. Some work has been done to evaluate the quality of the video summary by calculating the inclusion and redundancy automatically based on pre-defined ground truth [4, 12, 16]. But the uniform framework with comprehensive consideration for objective assessment 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.

To address the problem of current work on objective assessment for video summary, we propose a uniform framework based on the human's evaluation criteria for video summary and several novel algorithms to calculate these criteria automatically. The framework and algorithms mainly focus on full-reference quality assessment for video summary, meaning that the candidate video summary is evaluated based on the comparison with a pre-defined reference video summary.

Full-reference assessment is initially defined by Wang to evaluate the quality loss of the image after some processing via comparing with a complete perfect reference, such as the original image [15]. Relatively, there exist non-reference image assessment [11] and reduced-reference image assessment [6]. Full-reference video summary assessment is stemmed from full-reference image quality assessment. But different with image quality assessment, which is easy to achieve a consistent opinion of a good reference,

978-0-7695-3503-6/08 $25.00 © 2008 IEEE
DOI 10.1109/ICDMW.2008.55

people may have different perfect summaries in their brain under video summary assessment. So in this paper, we only work on the case that a unique and perfect reference summary is given. Non-reference and reduced-reference video summary assessment for different user's preferences will be considered in the future work.

The paper is organized as follows: Section 2 introduces current assessment methods for video summary. Section 3 proposes a novel framework for full-reference video summary assessment and provides several algorithms to implement automatic quality assessment under this novel framework. Section 4 shows the performance of the proposed framework and techniques by experimenting with the standard datasets. The paper is closed with conclusion and further work.

2. Related work

Based on the difference of human's interaction, current quality evaluation methods for video summary can be further categorized to subjective evaluation and objective evaluation [14]. Subjective evaluation mainly involves independent users judging the quality of the generated video summaries and calculates the cognitive value based on psychological metrics [4]. The direct and the most widely used subjective 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 [5]. But only using the overall score is too rough to describe the quality of the video summary. So different subjective measures are proposed to define the desirable characters for a good summary. A typical set of subjective measures was proposed by He et al. in [2], which provided the 4C criteria for an ideal summary:

-*Coverage*: the set of segments selected for the summary should cover all the "key" points.

-*Conciseness*: any segment of the talk that is 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 subjective assessment can be mainly recapitulated by the criterions or combinations of the criterions under these 4Cs criteria. For example, in the task of rushes summarization for TRECVID 2007, the criterion of ground-truth inclusion actually can be considered as one way to measure the coverage of the summary.

Although subjective evaluation is probably the most useful and realistic form of video summary evaluation [14], it suffers from human factors [10], high labor cost [14] and unreusable characters [3]. To address these problems, the objective assessment techniques for video summary are highly demanded. Currently, objective evaluation techniques can be classified into two categories. One category focuses on assessing the objective measure, such as the length of the summary [10], while another category works on providing the human like assessment by quantitative analysis of multimedia content. To map human's judgment, most objective methods manually define a set of ground-truth or/and keyframes. Silva et al. [12] and Yahiaoui et al. [16] 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. [4] calculates precision, recall and redundancy rate by matching the pre-defined ground truths in order to evaluate the content coverage and redundancy of video summary. Unfortunately, a uniform framework with comprehensive considerations of objective assessment for video summary is unavailable yet.

3. Video summary quality assessment

Figure 1 shows the framework of full-reference quality assessment system for video summary. The reference video summary is generated manually and assumed to be the only perfect abstraction of the original video file. We don't consider that one video file has alternative perfect summaries in this paper. Therefore, the full-reference quality assessment of video summary is formalized to the problem of pair-wise video sequence comparison for assessment purpose.

We first partition the reference summary and the candidate summary into a set of Summary Units (SUs) respectively. Then we match the SUs in the two sequences using algorithm of alignment based summary unit matching. After it, we calculate the quality of the candidate summary individually from four aspects: coverage, conciseness, coherence, and context, which are consistent with subjective criteria of an ideal summary [2]. To map human's judgment, we explore learning based weight adaptation method to integrate the quality scores of 4C assessment.

3.1. Summary unit sequence generation

Simply speaking, video summary can be described as a component sequence with the appropriate order. We define the component of the video summary as "summary unit". SU can be a video scene, shot, sub-shot and even a frame for different video files and different summarization targets. Definitely, if the spatial separability is permitted, SU can be a special location 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

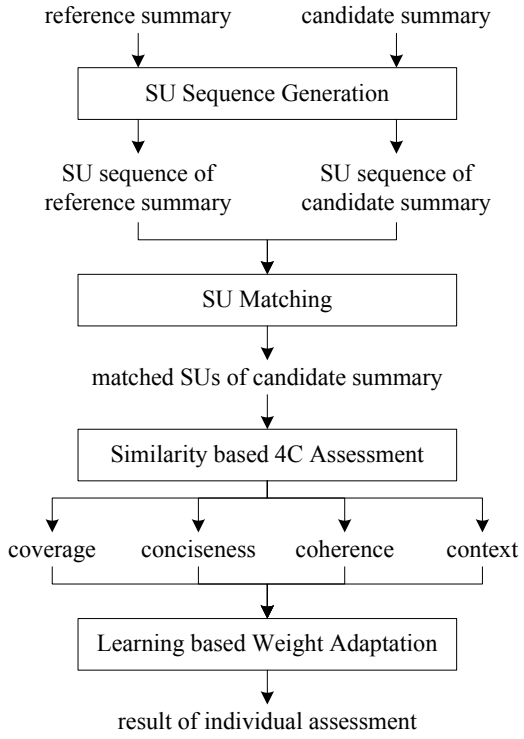

Figure 1. Framework of full-reference quality assessment system of video summary.

close caption. Due to the page limitation, we only consider the temporal separability of the video file for video summary quality assessment, i.e. subshot is used in this paper.

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

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

Where, M, N are the SU numbers of RS and CS respectively. The following assessment is based on the comparison of the two SU sequences. In this paper, we generate the SU sequences using the twin-comparison algorithm [18] on the HSV feature space of local color histogram.

3.2. Alignment based summary unit matching

In this part, we check every SU in the candidate summary and look for the most similar one in the reference summary. Various algorithms are available for subshots matching [17]. Regarding the SU as a time-order frame sequence, we use a well-known Needleman-Wunsch algorithm [9] in this paper.

We define SU_{R_i} in the reference summary as a frame sequence $\{f_{R_i1}, f_{R_i2}, ...\}_m$ and SU_{C_j} in the candidate summary as a frame sequence $\{f_{C_j1}, f_{C_j2}, ...\}_n$, here m and n are the frame numbers of SU_{R_i} and SU_{C_j} respectively. A feature vector v is extracted for each frame in SU_{R_i} and SU_{C_j}, and the Euclidean distance in the feature space $dis(f_{R_ip}, f_{C_jq})$ between frames f_{R_ip} and f_{C_jq} is used to determine if f_{R_ip} matches f_{C_jq} with a pre-defined threshold thr_{fs}. If $dis(f_{R_ip}, f_{C_jq}) < thr_{fs}$, f_{R_ip} can match f_{C_jq} and their similarity is:

$$Sim(f_{R_ip}, f_{C_jq}) = 1 - dis(f_{R_ip}, f_{C_jq}). \tag{2}$$

Next, we use Needleman-Wunsch algorithm to achieve the optimal matching of SU_{R_i} and SU_{C_j}. Needleman-Wunsch algorithm utilizes dynamic programming in matching and the objective function in alignment is defined as follows:

$$\begin{aligned} s_{p,1} &= Sim(f_{R_ik}, f_{C_j1}) \\ s_{1,q} &= Sim(f_{R_i1}, f_{C_jq}) \\ s_{p,q} &= max(s_{p,q-1}, s_{p-1,q}, s_{p-1,q-1} + Sim(f_{R_ip}, f_{C_jq})) \end{aligned} \tag{3}$$

Here, $s_{m,n}$ can be treated as the score of SU_{R_i} and SU_{C_j} alignment. Considering SU_{R_i} and SU_{C_j} may partly match, that means SU_{C_j} may lose some frames of SU_{R_i} or contain some redundant frames, we calculate the final alignment score as follows:

$$align(SU_{R_i}, SU_{C_j}) = \frac{1}{min(m,n)} s_{m,n}. \tag{4}$$

If the maximal matching score for SU_{C_j}, according to some SU_{R_i} in all the SUs in the reference summary, is higher than the pre-defined threshold thr_{SUs}, SU_{C_j} is considered to match SU_{R_i}; otherwise, SU_{C_j} is considered as a noise. The summary unit matching algorithm is provided in Table 1.

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

3.3. Similarity based 4C assessment

In this part, we assess the scores of 4C criteria using the result of SU matching. In 4C assessment, the influence caused by the same problem are avoided to be repetitive calculated in different aspects. For example, if two SUs are inverted in the candidate summary, we only consider the influence to SU order in context but ignore the influences in other aspects. In the following, we discuss the assessment of coverage, conciseness, coherence and context respectively.

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

Algorithm: Summary unit matching
Input:
$SU_{C_j} = \{f_{C_j1}, f_{C_j2}, ...\}_n$
$SU_{R_k} = \{f_{R_k1}, f_{R_k2}, ...\}_m, \forall k, k \in \{1, 2, ..., M\}$
Output:
SU_{R_i} or NULL
1. for each $SU_{R_k} = \{f_{R_k1}, f_{R_k2}, ...\}_m \in RS$
$$score_{R_k} = align(SU_{R_k}, SU_{C_j})$$
2. select $SU_{R_i} \in RS$ with the maximal score:
$$i = \arg \max_{1 \le k \le M} (score_{R_k})$$
3. if $score_{R_i} > thr_{SUs}$, return $score_{R_i}$ else, return NULL
Where,
RS: reference summary
SU_{R_k}: any SU in the reference summary
SU_{C_j}: a SU in the candidate summary
f_{R_kp}: any frame in SU_{R_k}
f_{C_jq}: any frame in SU_{C_j}

3.3.1 Coverage assessment

Coverage of the candidate summary is 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 coverage of all the SUs in the reference summary:

$$Cov(CS) = \sum_{i=1}^{M} Cov(SU_{R_i}). \quad (5)$$

To 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 which match SU_{R_i}, we choose the best SU_{C_j} to calculate.

The covered content of SU_{R_i} can be calculated as the sum of the covered content of the frames in S_{R_i} based on the result of SU matching. Figure 2 shows an example of the result of SU matching. If a frame in S_{R_i} matches a corresponding frame in SU_{C_j}, we call it "matched frame"; otherwise, we call it "unmatched frame". To a matched frame, such as f_{R_i5}, its covered content can be calculated as the similarity between it and its corresponding frame. To an unmatched frame, such as f_{R_i3}, its content may be partly covered by the corresponding frames of its nearest matched frame, for the adjacent frames in a video file are usually interrelated in content which calls "temporal redundancy" of video characteristics.

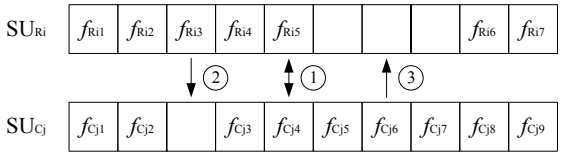

Figure 2. Aligned frame sequences of two summary units in the reference summary and the candidate summary.

To clearly explain the covered content calculation of SU_{R_i}, we define a concept "related frame". To a matched frame, its related frame is the corresponding frame which matches it. To 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 most similar corresponding frame to current unmatched frame as its related frame. For example, in Figure 2, f_{R_i5} matches f_{C_j4} and the related frame of f_{R_i5} is f_{C_j4}. f_{R_i3} does not match any frame in SU_{C_j}, so we look for the nearest matched frame(s) of f_{R_i3} (f_{R_i2} and f_{R_i4}) in SU_{R_i} and select the more similar corresponding frame from f_{C_j2} and f_{C_j4} as the related frame of f_{R_i3}. The coverage of SU_{R_i} can be calculated as follows:

$$Cov(SU_{R_i})_{SU_{C_j}} = \sum_{p=1}^{m} Sim(f_{R_ip}, RF(f_{R_ip})). \quad (6)$$

Where, $RF(f_{R_ip})$ is the related frame of f_{R_ip} in SU_{C_j}.

3.3.2 Conciseness assessment

Conciseness of the candidate summary is 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:

$$Coc(CS) = \sum_{j=1}^{N} Coc(SU_{C_j}). \quad (7)$$

To 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, $Coc(SU_{C_j})$ is 0; if only SU_{C_j} but no other SUs 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 choose the best SU_{C_j} to calculate and consider the conciseness 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 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:

$$Coc(SU_{C_j})_{SU_{R_i}} = \sum_{q=1}^{n} Sim(f_{C_j q}, RF(f_{C_j q})). \quad (8)$$

Where, $RF(f_{C_j q})$ is the related frame of $f_{C_j q}$ in SU_{R_i}.

3.3.3 Coherence assessment

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

We consider the coherence of the candidate summary in two aspects, inner SU coherence and inter SU coherence, and define the coherence of the candidate summary as follows:

$$Coh(CS) = w_1 * Coh_{inner}(CS) + w_2 * Coh_{inter}(CS). \quad (9)$$

Where, w_1 and w_2 are weight parameters.

We define the inner SU coherence of the candidate summary as the sum of the inner coherence of each SU:

$$Coh_{inner}(CS) = \sum_{j=1}^{N} Coh_{inner}(SU_{C_j}). \quad (10)$$

The inner coherence of SU_{C_j} is calculated by comparing to its corresponding SU_{R_i} in the reference summary. The noise SUs are ignored in inner coherence evaluation and their inner coherence are considered to be 0.

To calculate the inner coherence of each SU, we define "average distance" between two frames f_p and f_q as follows:

$$Avgdis(f_p, f_q) = \begin{cases} \frac{1}{q-p}\sum_{k=p}^{q-1} dis(f_k, f_{k+1}), & p < q \\ 0, & p \geq q \end{cases} \quad (11)$$

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:

$$Coh_{inner}(SU_{C_j})_{SU_{R_i}} = K_1 - \sum_{k=1}^{n-1} \max\Big(0, dis(f_{C_j k}, f_{C_j(k+1)}) - Avgdis(RF(f_{C_j k}), RF(f_{C_j(k+1)}))\Big). \quad (12)$$

Where, $RF(f_{C_j k})$ is the related frame of $f_{C_j k}$ in SU_{R_i}, K_1 is a positive constant to keep the result nonnegative.

We assess the inter SU coherence by comparing the mean distance values between two adjacent SUs in the reference summary and the candidate summary. The distance between two adjacent SUs is calculated as the distance between the last frame of the fore SU and the first frame of the latter SU. The inter SU coherence is calculated as follows:

$$Coh_{inter}(CS) = K_2 - \max\Big(0, \frac{\sum_{j=1}^{N-1} Dis(SU_{C_j}, SU_{C_{(j+1)}})}{N-1} - \frac{\sum_{i=1}^{M-1} Dis(SU_{R_i}, SU_{R_{(i+1)}})}{M-1}\Big). \quad (13)$$

Where, K_2 is a positive constant to keep the result nonnegative.

3.3.4 Context assessment

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

According to the assessment principle provided in the beginning of section 3.3, we ignore the influence to context caused by repeating or missing some SUs in the candidate summary. If a SU in the candidate summary is a noise, we ignore it in the context assessment; if more than one SUs match the same SU in the reference summary, we retain one of the SUs in context assessment each time and compute the mean value of the context scores in all situations. So the context of the candidate summary is defined as follows:

$$Cot(CS) = \frac{1}{N_s}\sum_{k=1}^{N_s} Cot_k(CS)$$
$$N_s = \prod_{i=1}^{M} \max(1, n_i) \quad (14)$$

Where, n_i is the number of SUs in the candidates summary match SU_{R_i} in SU matching and N_s is the number of all possible situations.

To calculate the context score, we define the order of SUs. Assume SU_i and SU_j are two SUs in a video sequence S, we define $O_S(SU_i, SU_j)$ as the order of SU_i and

SU_j in S: if SU_i appears before SU_j in S, $O_S(SU_i, SU_j)$ equals 1; otherwise, $O_S(SU_i, SU_j)$ equals 0.

To SU_{C_j} and SU_{C_q} in the candidate summary, we define "inversion" as follows: Assume SU_{C_j} and SU_{C_q} match SU_{R_i} and SU_{R_p} in SU matching respectively, if $O_{CS}(SU_{C_j}, SU_{C_q}) \neq O_{RS}(SU_{R_i}, SU_{R_p})$, $Inv(SU_{C_j}, SU_{C_q})$ equals 1; otherwise, $Inv(SU_{C_j}, SU_{C_q})$ equals 0.

We define the context of the candidate summary as follows:

$$Cot_k(CS) = \sum E(SU_{C_j}, SU_{C_q}) * Inv(SU_{C_j}, SU_{C_q}). \quad (15)$$

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

In this paper, we assume 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 is calculated as follows:

$$E(SU_{C_j}, SU_{C_q}) = \begin{cases} F(|p-i|), & O_{RS}(SU_{R_i}, SU_{R_p}) = 1 \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

Where, F is a decreasing function, i.e., $F(x) = 1/x$.

The above methods are integrated into the similarity based 4C assessment algorithm. We normalize the scores to the range of [0,1], and higher score means better quality in the corresponding aspect.

3.4. Learning based weight adaptation for assessment

In real applications, the user may require individual assessment with other criteria. For example, TRECVID 2007 requires assessing the video summary in ground truth inclusion, ease of understanding and lack of redundancy.

We assume the assessment results with other criteria can be acquired by combining the scores of 4C assessment. In this paper, we provide a linear combination approach by learning the weights of the 4C scores.

We combine the four score of 4C assessment and a constant element to a $1 * 5$ vector C:

$$C = (c_1, c_2, c_3, c_4, c_5) = (K, cov, coc, coh, cot). \quad (17)$$

Where, cov, coc, coh, cot are the scores of 4C assessment respectively. K is a positive constant, i.e. $K = 1$.

We assume the required assessment criteria including n aspects and the scores of all the aspects can be represented as a $1 * n$ vector S:

$$S = (s_1, s_2, ..., s_n). \quad (18)$$

Table 2. Learning based weight adaptation for individual evaluation.

Algorithm: Learning based weight adaptation
Input:
cov, coc, coh, cot
label data S by subjective evaluation
Output:
weight matrix W
1. formulize the problem as follows: $C * W = S$, $C = (K, cov, coc, coh, cot)$
2. calculate W by the least squares method with the objective function: $W = \arg\min \sum \left\| \sum_{i=1}^{5} (c_i * w_{ij}) - s_j \right\|^2$
Where, cov: coverage score; coc: conciseness score; coh: coherence score; cot: context score; S: subjective evaluation result

Then, the transform between C and S can be represented as:

$$C * W = S. \quad (19)$$

Where, W is a $5 * n$ matrix.

We calculate the weight matrix W using the least squares method on training data as shown in Table 2. Based on the weight matrix W, we can calculate the vector S on the new dataset.

4. Experiments

We validate the performance of the proposed full-reference assessment techniques for video summary on the standard dataset from TRECVID 2007 rushes summarization task.

In the dataset, the rushes videos, which are the unedited raw footages with considerable noise and redundancy, are provided as the original video files. The reference summary is generated manually by assembling the frames extracted from these video files. The video summaries generated by different participates of TRECVID are considered as candidate summaries. In this section, the first experiment provides the procedure and the result of automatic assessment on one video file. The second experiment shows that how automatic assessment maps human's judgment well on the large-scale dataset.

4.1. Experiment on one video shot

We demo our proposed techniques 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 first partition the reference summary and the candidate summary to a set of SUs as shown in Figure 3. The feature used in SU sequence generation and SU matching is the local color histogram. Each frame of video summary is divided into 4*4 sub-images of the same sizes and shapes. For each sub-image, 16 bins color histogram of HSV color model is extracted according to MPEG-7. Then each frame can be represented with 256-bins feature vector. The distance of two feature vectors is Euclidean distance in our experiment.

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

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

Table 4 shows the quality 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 artificial summaries with the obvious problems in coverage, conciseness, coherence and context respectively. Candidate summary 2 misses the last two SUs of reference summary, 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 participate. These two candidate summaries are generated by multimedia content analysis algorithms. Obviously, their performances are not as good as the artificial dataset and the problems of quality are more complicated.

Table 4. 4C assessment results on shot 103 in MRS044500.

CS No.	cov	coc	coh	cot
#1	1.000	1.000	1.000	1.000
#2	0.750	1.000	1.000	1.000
#3	1.000	0.800	1.000	1.000
#4	0.954	1.000	0.889	1.000
#5	1.000	1.000	1.000	0.741
#6	0.904	0.906	0.903	1.000
#7	0.690	0.421	0.827	0.623
#8	0.794	0.582	0.828	0.873

4.2. Experiment on large-scale dataset

In this section, we will show how to integrate the 4Cs criteria to adapt different evaluation requirements. In TRECVID 2007, each candidate summary is assessed manually from seven aspects. Four of them are objective measures, such as the length of the summary. The left three are subjective measures of "INclusion of ground-truth" (IN), "EAse of understanding" (EA), and "lack of REdundancy" (RE) [10]. We will calculate these three subjective measures using the proposed methods and compare with human's judgment.

We buildup the test-bed using ten participants and ten video files. The selected participants are: attlabs, cityu, cmu, cost292, hkpu, kddietal, ntu, thu-icrc, ucal, umadrid. These ten groups with different performances in the competition are selected from total twenty-four participants. The ten video files only include one story of multiple retakes for each file. They are one part of these rushes: MRS025913, MRS042543, MRS042548, MRS043400, MRS044500, MRS048779, MRS145918, MRS157445, MRS157475, MS210470.

For each video file in our experiment, ten participants have ten candidate summaries, so totally there are one hundred candidate summaries for ten video files. For these summaries, we calculate 4Cs score for all of them using similarity based 4Cs assessment.

In the learning based weight adaptation part, we train and test using the similar experimental environment of TRECVID 2007. Three assessors are used to grade the IN, EA, and RE of these candidate summaries following the procedures in [10]. We randomly select 60% video summaries of each original video file as the training data and

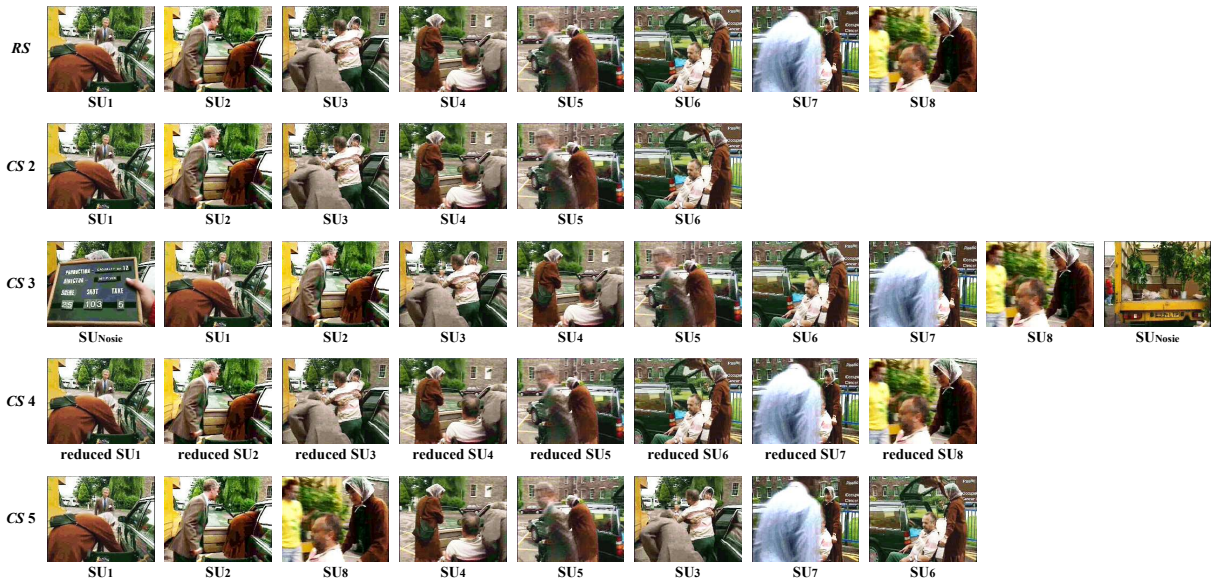

Figure 3. Video summaries for the rushes file of shot 103 in MRS044500.

Table 5. Result of learning based weight adaptation.

w_{ij}	K	cov	coc	coh	cot
IN	-0.079	0.945	0.135	0.006	0.039
EA	-2.304	3.708	0.839	0.930	2.189
RE	-2.013	0.958	4.159	0.809	1.224

utilize the rest 40% video summaries as the test data. Table 5 gives the weight values after training. Figure 3 demos the correlation between 4C and IN, EA, RE. For example, coverage domains the IN while the other three criteria also have some influence on human's judgment of inclusion. Figure 5 shows the correlation between subjective grading and objective assessment on the test data. It is obvious that the objective assessment techniques proposed by us can map human's evaluation very well.

5. Conclusions

This paper presents a novel framework to assess the quality of the video summary against the given reference. The framework replies on three underlying algorithms that are well-adapted to the characteristics of video summary assessment: alignment based summary unit matching, similarity based 4C assessment, and learning based weight adaptation. Together, they provide a complete objective assessment framework that well maps the subjective evaluation by human being. 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 assessment method without the requirement of a perfect reference summary, i.e., non-reference or reduced-reference assessment for video summary. Second, current weight adaptation algorithm is based on the assumption of linear combination model of 4C criteria. We will consider the possibility of other models and compare the assessment performance with linear model.

6. Acknowledgments

This work is supported by grant G-YG28 Central Research Grant and the National Natural Science Foundation of China (60533080).

References

[1] Y. Gong and X. Liu. Summarizing video by minimizing visual content redundancies. *Int'l Conf. Multimedia and Expo*, 2001.

[2] L. He, E. Sanocki, A. Gupta, and J. Grudin. Auto-summarization of audio-video presentations. *ACM Int'l Conf. Multimedia*, 1999.

[3] X. S. Hua, W. Liu, and H. J. Zhang. An automatic performance evaluation protocol for video text detection algorithms. *IEEE Trans. Circuits and Systems for Video Technology*, 14(4):498–507, 2004.

[4] M. Huang, A. B. Mahajan, and D. F. DeMenthon. Automatic performance evaluation for video summarization. *Tech. Report of Maryland University*, 2004.

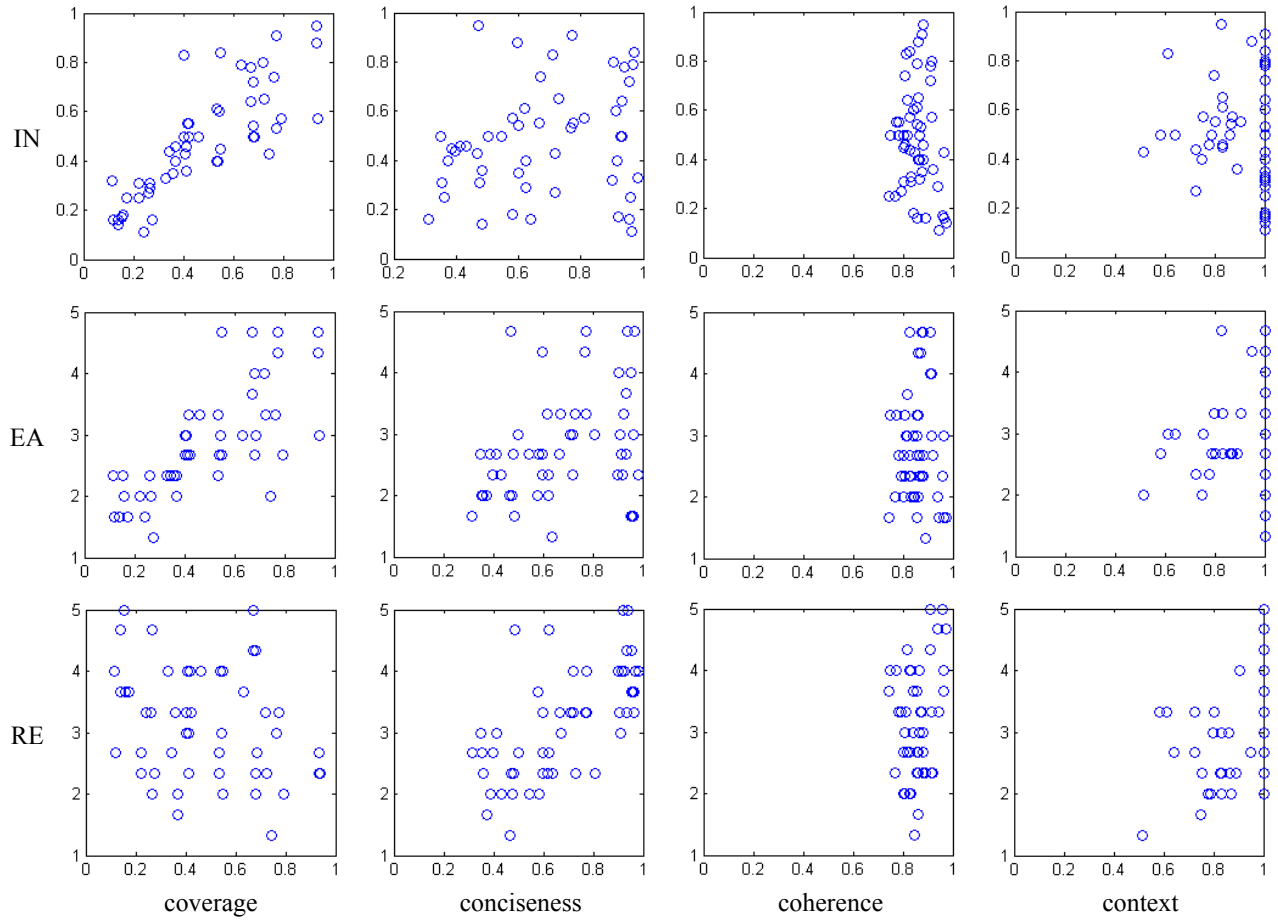

Figure 4. Correlation between 4C assessment results and IN, EA, RE scores on the training data.

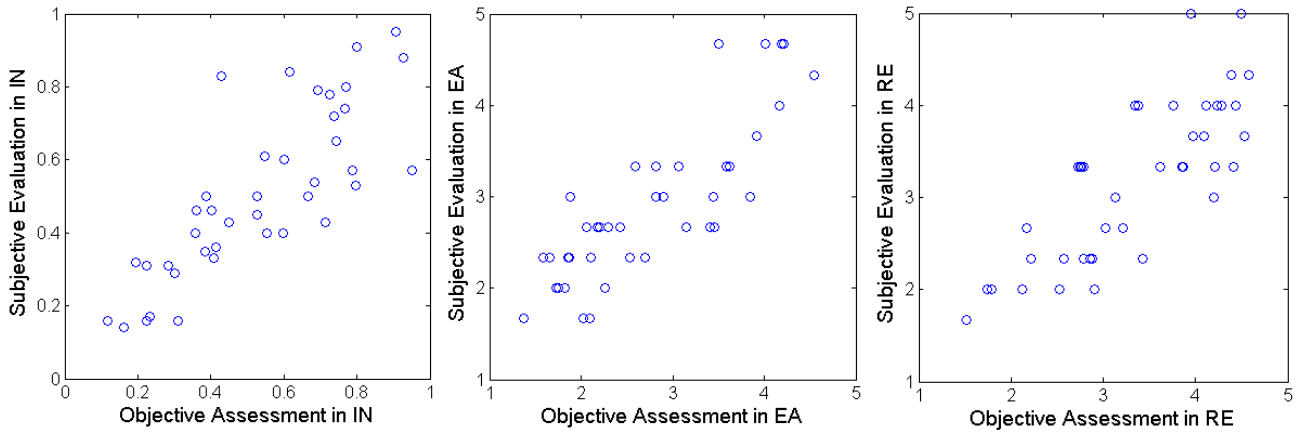

Figure 5. Correlation between subjective evaluation and objective assessment on the test data.

[5] H. Knoche, H. G. D. Meer, and D. Kirsh. Utility curves: Mean opinion scores considered biased. *Int'l Workshop Quality of Service*, 1999.

[6] T. M. Kusuma and H. J. Zepernick. A reduced-reference perceptual quality metric for in-service image quality assessment. *Mobile Future and Symp. Trends in Communications*, 2003.

[7] Y. F. Ma, L. Lu, H. J. Zhang, and M. Li. A user attention model for video summarization. *ACM Int'l Conf. Multimedia*, 2002.

[8] H. R. Naphide and T. S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Trans. Multimedia*, 3(1):141–151, 2001.

[9] S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. molecular biology*, 48(3):443–453, 1970.

[10] P. Over, A. F. Smeaton, and P. Kelly. The trecvid 2007 bbc rushes summarization evaluation pilot. *Int'l Workshop TRECVID Video Summarization*, 2007.

[11] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. Blind quality assessment for jpeg2000 compressed images. *Conf. Signals, Systems and Computers*, 2002.

[12] G. C. Silva, T. Yamasaki, and K. Aizawa. Evaluation of video summarization for a large number of cameras in ubiquitous home. *ACM Int'l Conf. Multimedia*, 2005.

[13] C. M. Taskiran. Evaluation of automatic video summarization systems. *Conf. Multimedia Content Analysis, Management, and Retrieval*, 2006.

[14] B. Truong and V. S. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Computing, Communication, and Application*, 3(1):1–37, 2007.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004.

[16] I. Yahiaoui, B. Merialdo, and B. Huet. Comparison of multiepisode video summarization algorithms. *EURASIP J. Applied Signal Processing*, 2003(1):48–55, 2003.

[17] M. M. Yeung and B. Liu. Efficient matching and clustering of video shots. *Int'l Conf. Image Processing*, 1995.

[18] H. J. Zhang, A. Kankanhalli, and S. W. Smoliar1. *Readings in Multimedia Computing and Networking*, chapter Automatic partitioning of full-motion video, pages 321–339. Morgan Kaufmann, 2001.