Evaluation of Video Summaries

Yingbo Li and Bernard Merialdo
Institut Eurecom, France
{Yingbo.Li, Bernard.Merialdo}@eurecom.fr

Abstract

In this paper we present a demonstration tool to illustrate the VERT measure for evaluating video summaries. An interface allows the user to conveniently watch the videos, and select the keyframes for his summary. Then a panel of colored grids shows the comparison of his selection with the reference, and another panel shows the VERT weights for the user selection and the reference. Finally, curves of the VERT scores are displayed in the last panel. In this paper, we review the basic principles of the VERT measure and then describe the demo tool in more details.

1. Introduction

With the popularization of the digital video, common people can easily create their own videos. The amount of these personal videos, and professional videos, such as news, sports, advertisements and MV, is increasing day by day. Therefore, searching among large quantities of videos is a focused topic in the domain of multimedia processing, of which video summarization is a key component. A lot of methods have been proposed [6] [7]. Some algorithms are domain-specific, like on sports or music, while the others are more generic. Some exploit only the image information inside the video; others only use the audio information while a few use both. Some algorithms even exploit external data such as motion and position of the video camera. But a major issue in video summarization remains the evaluation, which allows comparing and improving algorithms. We have already proposed the VERT (Video Evaluation by Relevant Threshold) measure [4] for the evaluation of video summaries. The VERT measure is inspired from the ROUGE measure already developed for textual summaries. It formalizes the comparison between a candidate summary and a set of human-generated reference summaries and has the potential of becoming a standard in the field. In this paper, we expose the VERT measure and describe a demo tool to visualize the evaluation process.

The paper is organized as follow: In Section 2 we review the basic principle and evolution history of VERT. Then in Section 3 we explain in detail our demo tool with its five tabs. The last section is the conclusion.

2. The VERT Meaure

The VERT algorithm used in the evaluation of video summarization is inspired by the BLEU algorithm in the evaluation of machine translation, and the ROUGE algorithm in the evaluation of text summarization.

2.1. BLEU

The BiLingual Evaluation Understudy (BLEU) [1] is the earliest and fundamental algorithm. BLEU based on n-gram co-occurrence scoring, has been used to automatically evaluate machine translations. It is now the scoring metric used in the NIST (NIST 2002) translation benchmarks. BLEU compares a candidate translation with several human-generated reference translations using n-gram co-occurrence statistics. The results of the BLEU measure have been shown to have a high correlation with human assessments. BLEU is a precision metric, defined by the following formula:

$$BLEU_n = \frac{\sum_{C\in\{CandidateSentences\}}\sum_{gram_n\in C} Count_{clip}(gram_n)}{\sum_{C\in\{CandidateSentences\}}\sum_{gram_n\in C} Count(gram_n)} \quad (1)$$

where $Count_{clip}(gram_n)$ is the maximum number of n-grams co-occurring in the candidate translation and one of the reference translations, and $Count(gram_n)$ is the number of n-grams in the candidate translation. The computation is performed sentence by sentence.

2.1. ROUGE

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure proposed by Lin [2] [3] has been proved to be a successful algorithm to complete the task of evaluating the text summarization. This measure counts the number of overlapping units between the summary candidates generated by computer and several ground truth summaries built by humans. In [3], several variants of the measure are introduced, such as ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S. Because VERT reuses the ideas of ROUGE-N and ROUGE-S, we briefly review only these two. ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. It is defined by the following formula:

$$ROUGE - N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (2)$$

where n is the length of the n-gram, gram_n , $Count(gram_n)$ is the number of n-grams in the reference summaries, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. ROUGE-S is Skip-Bigram Co-occurrence Statistics. And Skip-bigram is any pair of words in their sentence order, allowing for arbitrary gaps. Skip-bigram co-occurrence statistics measure the overlap of skip-bigrams between a candidate translation and a set of reference translations. To reduce the spurious matches, the maximum skip distance is limited in two in-order words that are allowed to form a skip-bigram.

2.1. VERT [4]

After ROUGE and BLEU, the algorithm of VERT was proposed. Compared with machine translation and text summarization, the video and its summarization owns their special properties: (1) These sequences of videos are segmented into shots or subshots, and eventually each shot is represented by one or more keyframes, (2) A selection of the video content to be included in the summary is performed. Eventually, this selection is ordered with the selection order. (3) The selected contents consist of a video summary. Depending on the intended format, the video summary may be a concatenated video, or a set of keyframes with specific presentation.

In VERT, the important keyframes are selected. Each keyframe is assigned a weight $w_S(f)$ depending on the rank of keyframe *f* in the selection *S*. Therefore, VERT measure compares a set of computer-selected keyframes with several reference sets of human-selected keyframes. By similarity with ROUGE-N, we first propose the VERT-N which is defined as:

$$VERT\text{-}N(C) = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} W_C(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} W_S(gram_n)} \quad (3)$$

where C is the candidate video summary, $gram_n$ is a group of n keyframes, $W_S(gram_n)$ is the weight of the group $gram_n$ for a reference summary S, and $W_C(gram_n)$ is the weight of the group $gram_n$ for the candidate summary C. We need to mention that in the numerator of the formula, the summation of $W_C(gram_n)$ is only taken for the gram_n which are present in the reference summary *S*. VERT-N is a recall-related measure and computes a percentage of gram_n from the reference summaries occurring in the candidate summary too.

The notion of “group of n keyframes” may be consecutive keyframes or the set of several non-consecutive keyframes. It is more sensible to define a “group of n keyframes” as a simple subset of size *n*, because sometimes the consecutive keyframes does not contain much more information than one keyframe in it. Therefore, the VERT-N resembles more to the ROUGE-S variant.

In [4], the study is restricted to the cases *n=1* and *n=2*. We thus define the VERT-1 and VERT-2 measures by Eq. 4 and Eq. 5:

$$VERT\text{-}1(C) = \frac{\sum_{S \in R} \sum_{f \in S} W_C(f)}{\sum_{S \in R} \sum_{f \in S} W_S(f)} \quad (4)$$

$$VERT\text{-}2(C) = \frac{\sum_{S \in R} \sum_{(f,g) \in S} W_C(f,g)}{\sum_{S \in R} \sum_{(f,g) \in S} W_S(f,g)} \quad (5)$$

And VERT-2 owns two variants:

1) VERT-2_S: $W_S(f,g) = \frac{w_S(f)+w_S(g)}{2}$
2) VERT-2_D: $W_S(f,g) = |w_S(f) - w_S(g)|$

According to the experiments of [4], VERT-2_D is the best if the keyframes weights are different according to the selection order, and VERT-2_S is the best if the keyframes weights are uniform. In this paper, we use the VERT-2_D variant with non-uniform weights.

3. VERT Demonstration Tool

We downloaded a set of videos, “YSL”, from a news aggregator website (http://www.wikio.fr), as the experimental videos. This is similar with the data in [4]. This website gathers news items dealing with the same specific topic and originating from different sources. The “YSL” set contains 6 videos about the death of a famous designer. Some videos represent the burial, some display older fashions shows, while some are interviews or comments. It is possible that some videos are incorrectly classified and unrelated to the topic.

Our demonstration tool is designed as a tabbed dialog box. In this dialog, there are five tabs: the first one is “6 videos”; the second one is “keyframes selection”; the third one is “Selection show”; then “VERT-2_D weights of reference and user selection” tab is shown; the name of the last tab is “VERT curves”. These tabs are shown in Fig. 1. We will describe these five tabs in detail in the remaining part of this section.

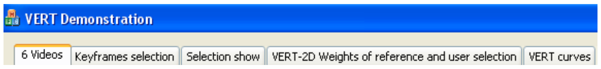

Figure 1. Tabs of VERT Demonstration

3.1. The tab “Videos”

In this tab, 6 videos in our prepared video set “YSL” are played in the demonstration tool.

1) “Video 1” is a long video of 7 minutes 38 seconds, whose content is the fashion show of the female clothes.
2) “Video 2” is a video with the length of 29 seconds. It is an advertisement of the perfume.
3) “Video 3” is an advertising for food. Its time length is 19 seconds.
4) “Video 4” with the length of 2 minutes 20 seconds is the news of CNN about a French clothes designer.
5) “Video 5” is a MV, which tell us something of skating, arms and so on. Video 5 owns the time of 3 minutes 12 seconds.
6) The last video “Video 6” is like an interview of a singer from Hong Kong by the brand “Cartier”. The length of the time is 3 minutes 6 seconds.

The design of this tab “Videos” is shown in the following figure, Fig. 2.

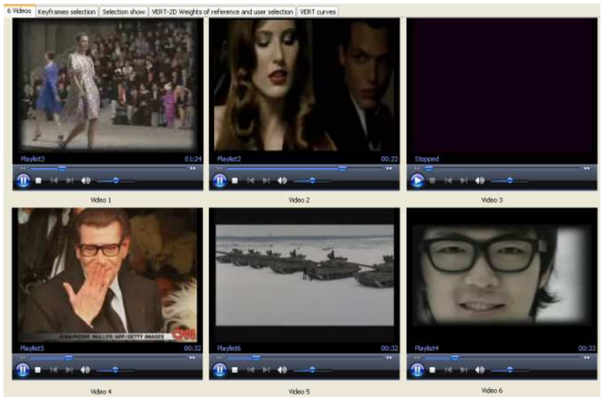

Figure 2. The tab “Videos”

3.2. The tab “keyframes selection”

This tab is composed of 47 keyframes with the shape of 6 rows and 10 columns. The keyframes are obtained by Video-MMR [5]. Each row is from a video, and the order of columns is as the time order. In Fig. 3, there is a check box under each keyframe. The user could select 10 keyframes by the check boxes and at last click the button “Finish selection”. The reason of selecting 10 keyframes is that we own a reference set of keyframes with the size 10 keyframes by 12 people. After the selection of keyframes, the demonstration tool will compute the demo images and curves of Subsection 3.3, 3.4 and 3.5.

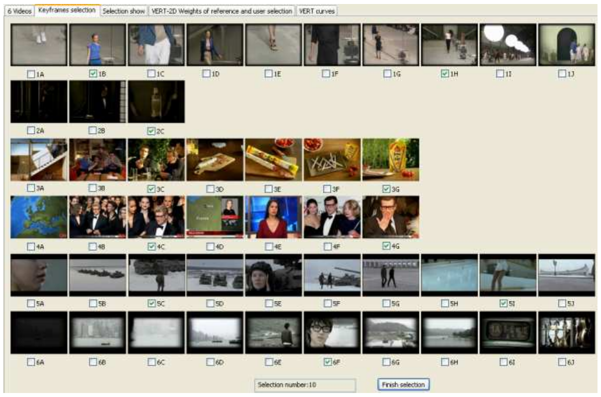

Figure 3. The tab “keyframes selection”

3.3. The tab “selection show”

When the user clicks the tab “selection show”, an image representing the selected keyframes by reference and user are shown in a 13x47 grid. The number 47 means the total 47 keyframes whose order is from the first video to the sixth video. Furthermore, the first row to the twelfth row is the selection by 12 people in the reference. The last row in the grid is just the user selection in the tab “keyframes selection”. The selection order is represented by the color. The color bar is shown right to the image in Figure 4. From this Fig. 4, we can intuitively read the user selection and reference selections of the keyframes.

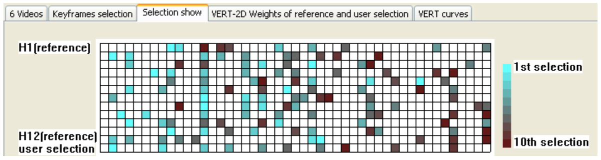

Figure 4. The tab “selection show”

3.4 The tab “VERT-2D weights of reference and user selection”

This tab contains two images of 47x47 grids. The number 47 owns the same meaning with the last subsection. Because VERT-2_D is 2-D computation and it considers the time order of two keyframes in a 2-gram, the grids are two dimensions. The left grid presents the reference weights of VERT-2_D which are the denominator of Eq. 3, while the right grid is the user weights of VERT-2_D being the numerator of Eq. 3. The color bar shows the color from high weight to low weight in Figure 5 too. We can directly view which weights of VERT-2_D are selected by the user.

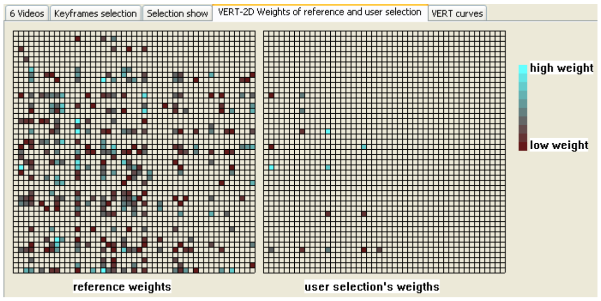

Figure 5. The tab "VERT-2D weights of reference and user selection"

3.4 The tab "VERT curves"

Fig. 6 shows two VERT-2_D curves, with 12 discrete points. The point "U, H1" means that it uses the 11 reference summaries excluding the first one as the new reference in Eq. 3, the first reference summary H1 as one candidate summary and user selection as another candidate summary. Green point of "U, H1" displays the VERT-2_D score of candidate summary H1 and Red point is the VERT-2_D score of candidate summary of user selection. The other points in Fig. 6 are similar. We can see that two curves are coherent, which means the reference is stable and trustable.

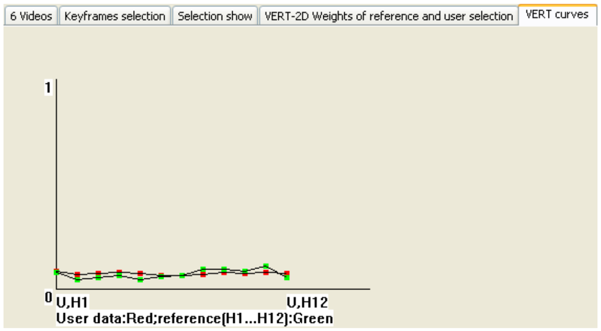
6 Videos
Keyframes selection
Selection show
VERT-2D Weights of reference and user selection
VERT curves
1
0
U,H1
U,H12
User data:Red;reference(H1...H12):Green

Figure 6. The tab "VERT curves"

4. Conclusions

We reviewed the VERT measure, which is used to evaluate video summaries. Based on this algorithm, we designed our demo tool to facilitate the procedure of user selection and the visualization of summary evaluation. With our demo tool, the user could easily analyze the similarity and difference between his selection of video keyframes and the reference. We can also see that the reference is coherent with the user selection. This tool provides an illustration of the mechanisms that underly the VERT measure.

This research is partially supported by the French ANR (Agence Nationale de la Recherche) in the project RPM2.

5. References

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a method for automatic evaluation of machine translation", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311-318, Philadelphia, July 2002.

[2] Chin-Yew Lin and Eduard Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics", *In Proceedings of the Human Technology Conference 2003*, Edmonton, Canada, May 27, 2003.

[3] Chin-Yew Lin, "ROUGE: a package for automatic evaluation of summaries", *In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, July 25 - 26, 2004.

[4] Yingbo Li, and Bernard Merialdo, "VERT: a method for automatic evaluation of video summaries", *Submitted.*

[5] Yingbo Li, and Bernard Merialdo, "Video summarization based on Video-MMR", *International Workshop on Image Analysis for Multimedia Interactive Services*, Italy 2010.

[6] Paul Over, Alan F. Smeaton, and Philip Kelly, "The trecvid 2007 bbc rushes summarization evaluation pilot", *ACM MM'07*, Augsburg, Bavaria, Germany, September 23–28, 2007.

[7] Arthur G.Money, "Video summarisation: a conceptual framework and survey of the state of the art", *Journal of Visual Communication and Image Representation*, Volume 19, 121-143, February 2008.