VERT: Automatic Evaluation of Video Summaries

Yingbo Li, Bernard Merialdo
EURECOM, Multimedia Communications Department
BP 193, 06904 Sophia Antipolis, FRANCE
{Yingbo.Li,Bernard.Merialdo}@eurecom.fr

ABSTRACT

Video Summarization has become an important tool for multimedia information processing, but the automatic evaluation of a video summarization system remains a challenge. A major issue is that an ideal "best" summary does not exist, although people can easily distinguish "good" from "bad" summaries. A similar situation arise in machine translation and text summarization, where specific automatic procedures, respectively BLEU and ROUGE, evaluate the quality of a candidate by comparing its local similarities with several human-generated references. These procedures are now routinely used in various benchmarks. In this paper, we extend this idea to the video domain and propose the VERT (Video Evaluation by Relevant Threshold) algorithm to automatically evaluate the quality of video summaries. VERT mimics the theories of BLEU and ROUGE, and counts the weighted number of overlapping selected units between the computer-generated video summary and several human-made references. Several variants of VERT are suggested and compared.

Categories and Subject Descriptors

I.2.10 [**ARTIFICIAL INTELLIGENCE**]: Vision and Scene Understanding – *video analysis.*

General Terms

Algorithms, Measurement, Human factor, Theory.

Keywords

Summary Evaluation, VERT, ROUGE, Video Summarization.

1. INTRODUCTION

The number of available videos is tremendously increasing daily. Some videos are from personal life, while others are recordings of TV channels, music clips, movies and so on. Therefore, video management has become an important research topic. Video summarization [4] [7] [11] is one of the key components for video management. A video summary [4] is a condensed version of the video information. It can provide the user with a fast understanding of the video content without spending the time to watch the entire video. The various forms of video summaries include: static keyframes, video skims and multidimensional browsers. Following single video summarization, multi-video summarization [5] [7] [9] has attracted many researchers recently. Multi-video summarization does not only need to consider the intra-relation among the keyframes in a single video, but also the inter-relation of the different videos in the same set. Consequently, the evaluation of video summaries [4] [6] [8] is a popular problem, still open to innovation. People can easily distinguish between "good" and "bad" summaries, but an ideal "best" summary does not exist, so that it is difficult to define a quality measure that can be automatically computed. It is still possible to set up experiments involving human beings to evaluate video summaries, but these experiments are costly, time-consuming, and cannot easily be repeated, which impairs the development of many algorithms based on machine learning techniques. A good quality measure achieving automatic computation, and showing a strong correlation with human evaluation is therefore of great interest.

Similar situations have already been encountered. In the domain of machine translation, BLEU [1] is a successful algorithm. The main idea of BLEU is to use a weighted average of variable length phrase matches against a set of reference translations. In the domain of automatic text summarization [10], ROUGE [2] [3] counts the n-grams of the candidate summaries co-occurring in the reference summaries to produce an automatic evaluation. In this paper, we propose VERT (Video Evaluation by Relevant Threshold), which uses ideas similar to BLEU and ROUGE, to automatically evaluate the quality of video summaries. It is suitable for both single and multi video. Red, green and blue being the three primary colors, ROUGE, VERT and BLEU, their French translations, could become the set of reference evaluation algorithms in their respective domains too.

This paper is organized as follows: Section 2 reviews BLEU and ROUGE, and Section 3 proposes VERT, together with its variants. Section 4 explains how the reference summaries are constructed and experimentally compares the variants of VERT. Finally this paper is concluded in Section 5.

2. RELEVANT KNOWLEDGE

2.1 BLEU

For automatically evaluating the quality of machine translation, the BiLingual Evaluation Understudy (BLEU) [1], based on n-gram co-occurrence scoring, has been proposed. It is now the scoring metric used in the NIST (NIST 2002) translation benchmarks. The main idea of BLEU is to measure the similarities between a candidate translation and a set of reference translations. BLEU compares a candidate translation with several human-generated reference translations using n-gram co-occurrence statistics. BLEU is defined:

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

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

MM'10, October 25–29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-60558-933-6/10/10...$10.00.

sentence. The results of the BLEU measure have been shown to have a high correlation with human assessments.

2.2 ROUGE

For text summarization evaluation, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure proposed by Lin [2] [3] has proved to be a successful algorithm. 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 our work reuses the idea of ROUGE-N and ROUGE-S, we briefly review both. ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. It is defined by the following formula:

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

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 in reference summaries.

ROUGE-N is a recall-related measure, while BLEU is a precision-based measure [2]. ROUGE-S is a Skip-Bigram Co-occurrence Statistics, where a skip-bigram is any pair of words in their sentence order, allowing for arbitrary gaps.

3. VERT

By borrowing ideas from ROUGE and BLEU, we extend these measures to the domain of video summarization. We focus our approach on the selection of relevant keyframes, as a video skim can be easily constructed by concatenating video clips extracted around the selected keyframes. Since we believe that the temporal order of keyframes is not as important as the word order in a sentence, we rather use the keyframe importance rank in the selection. The process of video summarization is then formalized as follows:

- We consider a set of video sequences $V_1, V_2, \ldots, V_k$ related to a given topic,
- These sequences are segmented into shots or subshots, and each shot is represented by one or more keyframes,
- Based on shots, subshots or keyframes, a selection of the video content to be included in the summary is performed. Eventually, this selection may be ordered, with the most important content being selected first.
- The selected content is assembled into a video summary, either in the form of a photo album or a video skim.

After the selection, each keyframe is assigned an importance weight $w_S(f)$ depending on the rank of keyframe *f* in the selection *S*. Therefore, our VERT measure compares a set of computer-selected keyframes with several reference sets of human-selected keyframes. Since BLEU is precision measure and ROUGE is recall measure, we propose VERT-Precision (VERT-P) and VERT-Recall (VERT-R) respectively.

3.1 VERT-Precision

Mimicking BLEU algorithm, we propose the VERT-P measure. Assume that we have k reference summaries (human selected lists), each containing n keyframes. Each keyframe is assigned an importance weight $W_S(x,y)$ according to its position in the selection ($x = 1, \ldots, k$ and $y = 1, \ldots, n$). Non-selected keyframes are assigned a weight of zero. Similarly the candidate summary (computer selected list) contains m keyframes, and each keyframe i is assigned a weight $W_C(i)$. VERT-P measures the precision of the position of each candidate keyframe in comparison to the reference summaries. For each keyframe i in candidate summary, the maximum weight that was assigned in the reference summaries is $T_i = \max_x W_S(x, y_i)$, where x is a reference summary, and y_i is the position of keyframe i in x. VERT-P compares this maximum weight with the actual weight that keyframe i was assigned in the candidate summary. This results in the following definition:

$$VERT\text{-}P=\frac{\sum_{i=1}^{m}\min\,[W_C(i),T_i]}{\sum_{i=1}^{m}W_C(i)} \quad (3)$$

The value of VERT-P is always between zero and one. The maximum is obtained when every keyframe of the candidate was selected with a weight that is lower than at least one of the human selections. For example, if a candidate keyframe was never selected by any human, the value of the measure will be strictly lower than 1. In this way VERT-P is a precision-based measure.

3.2 VERT-Recall

By similarity with ROUGE-N, we propose VERT-R$_N$:

$$VERT\text{-}R_N(C)=\frac{\sum_{s\in\{ReferenceSummaries\}}\sum_{gram_n\in S}W_C(gram_n)}{\sum_{s\in\{ReferenceSummaries\}}\sum_{gram_n\in S}W_S(gram_n)} \quad (4)$$

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. Note 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-R$_N$ is a recall-related measure too. As ROUGE-R$_N$, it computes a percentage of $gram_n$ from the reference summaries occurring also in the candidate summary. While ROUGE uses the notion of “word matching”, VERT-R considers the notion of “keyframe similarity”, which may be interpreted in a very strict sense (selection of the same keyframe), but also in a more relaxed manner by introducing a similarity measure between keyframes.

When *n* is larger than 1, the notion of “group of *n* keyframes” may have several interpretations. Since the selected summaries are ranked lists of keyframes, it is possible to consider consecutive keyframes in these lists. However, we decided that it was more sensible to define a “group of *n* keyframes” as a simple subset of size *n*, because the proximity of keyframes in the selected lists does not bear as much information as the order of words in a sentence. In this regard, VERT-R$_N$ resembles more to ROUGE-S.

In this paper, we restrict our study to the cases *n*=1 and *n*=2. We thus define VERT-R$_1$ and VERT-R$_2$ measures by Eq. 5:

$$VERT\text{-}R_1(C)=\frac{\sum_{S\in R}\sum_{f\in S}W_C(f)}{\sum_{S\in R}\sum_{f\in S}W_S(f)}$$
$$VERT\text{-}R_2(C)=\frac{\sum_{S\in R}\sum_{(f,g)\in S}W_C(f,g)}{\sum_{S\in R}\sum_{(f,g)\in S}W_S(f,g)} \quad (5)$$

In VERT-R$_1$, each $gram_1$ contains only 1 keyframe, so that the number of $gram_1$ is just the number of keyframes, and the weight of a group is simply the weight of the keyframe. Note that the denominator in Eq. 5 is actually the product of the total number of keyframes in all reference summaries times the sum of all weights. It’s a one-dimension computation.

In VERT-R$_2$, there are 2 keyframes in each $gram_2$, so it requires a two-dimension computation. We propose two variants for VERT-R$_2$: (1) VERT-R$_{2S}$, where the weight of a $gram_2$ is the average of the weights of the keyframes: $W_S(f,g)=\frac{w_S(f)+w_S(g)}{2}$; (2) VERT-R$_{2D}$, where the weight of a $gram_2$ is the absolute difference between the weights: $W_S(f,g)=|w_S(f)-w_S(g)|$. Obviously, VERT-R$_{2D}$ should only be considered if weights are non-uniform.

4. EXPERIMENTAL RESULTS

For our experiments, we downloaded two sets of videos, "DATI" and "YSL", from a news aggregator website (http://www.wikio.fr). This website gathers news items dealing with the same specific topic from different sources. "DATI" includes 16 videos, while "YSL" has 14 videos. The "DATI" set is about a politician woman: most are from TV news, showing either the person herself, or the comments about her. The "YSL" set contains videos related to the death of a famous designer. Some videos represent the burial; some are interviews; and some replay older fashions shows. Also some videos are irrelevant to the topic. We use a video summarization algorithm that we have previously developed, Video-MMR [9], for the initial keyframe representation of the videos.

This section is organized as follows: Subsection 4.1 explains the method for constructing the references by human assessment, and two systems of weights are suggested: ranking weights and uniform weights; Subsection 4.2 explains the principle of VERT evaluation; and the evaluation results are presented in Subsection 4.3.

4.1 Reference Construction

We now detail how we organized the construction of human-selected summaries which would be used as references. Our concern was to design a process which would facilitate the selection as much as possible, despite the complexity of the task.

1) For each video set, we identify 6 representative videos. For this, we compute the mean distance between each video and all the others in the set. Then we select the 3 videos with the smallest means, as containing the core of the set, and the 3 videos with the highest means, as containing the most distinctive information from the set.
2) On these 6 videos, we perform shot boundary detection, and one representative keyframe per shot is selected.
3) If a video produces more than 10 keyframes, we select the 10 most important keyframes by the Video-MMR algorithm. The result is a set of at most 60 keyframes that is representative of the visual content of the video set.
4) From these 60 keyframes, we ask each user to select the 10 most important frames as reference summaries. The selection is ordered, with the most important frame being selected first. Users may watch the original video if desired, and they can also access the related textual information.

The summaries of video sets "DATI" and "YSL" for constructing the references are shown in Fig. 1 and Fig. 2. The images in the same row originate from the same video. We enrolled 12 users, members of other projects in the laboratory, to select their own best summaries of 10 keyframes. In the Ranking Weights scheme, the weights decrease linearly from 1.0 (for the most important frame) to 0.1 (for the least important). In the Uniform Weights scheme, the weights are all equal to 0.1.

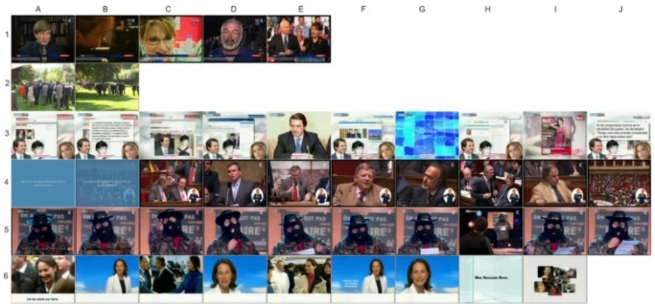

Figure 1. The set of keyframes of "DATI".

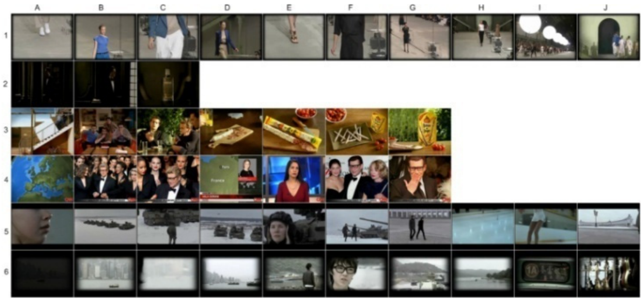

Figure 2. The set of keyframes of "YSL".

4.2 VERT Evaluation Principle

We want to evaluate if the values assigned by the VERT measures correlate with the human judgment on the quality of summaries. For each set "DATI" and "YSL", we constructed a set of 7 representative summaries of 10 keyframes each, including 2 random summaries, 1 summary constructed by K-Means, 2 summaries constructed by Video-MMR (with different parameter values), and the best and worst human summaries (based on our own judgment). From these 7 summaries, we created 21 pairs (an example from "YSL" is shown in Fig. 3) and asked humans to select the best summary in each pair. To reduce the load on users, the 21 pairs were separated into 2 groups and each group was evaluated by 6 users. In total, each pair has 6 evaluations (Human Pair Selection, HPS) from different humans identifying the best summary in the pair.

Figure 3. A Summary Pair of "YSL": One row, one summary.

4.3 VERT Evaluation

By applying VERT to the same 21 pairs, we can define the VERT Pair Selection (VPS), and compare it with the human selection (HPS). We use the accuracy percentage λ, and the Spearman rank correlation coefficient ρ [2,3] to quantify their correlation.

The Accuracy Percentage (AP) is the percentage of correct choices made by VPS compared with human reference, HPS. Let $P_i,\ i=1,\dots,21$ be the 21 pairs, we define:

$$C_{VERT}(i)=\begin{cases}-1 & \text{if the first summary is selected}\\+1 & \text{if the second summary is selected}\end{cases}$$

with a similar definition $C_{H_m}(i)$ for the choices of human m. The AP is defined as:

$$\lambda=\frac{1}{H}\sum_{m=1}^{H}\left[\frac{1}{21}\sum_{i=1}^{21}\frac{C_{VERT}(i)\cdot C_{H_m}(i)+1}{2}\right] \quad (6)$$

where $H = 6$ here. For the Spearman coefficient, we derive a ranking of the 7 summaries from VPS and HPS, and apply the formula:

$$\rho = 1 - \frac{1}{H}\sum_{m=1}^{H}\frac{6}{21(21^2-1)}\left[\sum_{i=1}^{21}\left(rank_{VERT}(i) - rank_{H_m}(i)\right)^2\right] \quad (7)$$

Table 1 and Table 2 show the APs and Spearman coefficients of VERT-P, VERT-R$_1$, VERT-R$_{2S}$ and VERT-R$_{2d}$ for the Ranking Weights system. We also evaluate human selection as the average of each HPS with the other 5 as references. We see that the VERT-R results are in the same range as human evaluation. For Uniform Weights, APs and Spearman coefficients are shown in Table 3, which does not contain VERT-P, because VERT-P is meaningless for Uniform Weights.

Table 1. λs with Ranking Weights

	P	R$_1$	R$_{2S}$	R$_{2d}$	User
DATI	0.5317	0.6270	0.5794	0.6270	0.5714
YSL	0.5317	0.7063	0.6905	0.6587	0.6286

Table 2. ρs with Ranking Weights

	P	R$_1$	R$_{2S}$	R$_{2d}$	User
DATI	0.1071	0.6429	0.4643	0.6429	0.6190
YSL	0.2143	0.7500	0.8571	0.8214	0.6310

Table 3. λs and ρs with Uniform Weights

	λ(R$_1$)	λ(R$_{2S}$)	ρ(R$_1$)	ρ(R$_{2S}$)
DATI	0.6270	0.5794	0.6429	0.4643
YSL	0.6905	0.6905	0.6071	0.8214

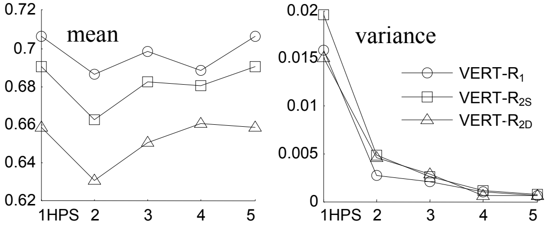

Figure 4. Means and Variances of λ for "YSL"

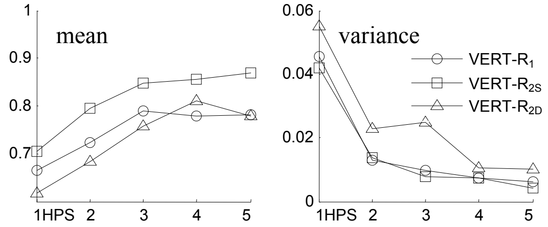

Figure 5. Means and Variances of ρ for "YSL"

In Fig. 4 and Fig. 5, we vary the number of HPS that are considered in the evaluation, from 1 to 5. We see that there is a convergence of the mean values λ and ρ, and that the variances of these values are greatly reduced when the full reference set is used. This is a clear indication that the values that have been computed are reliable estimates.

It is clear that the values of Spearman coefficients for VERT-P are very small, which indicates that VERT-P does not match well with human assessment. For VERT-R, the value of APs and Spearman coefficients are both around 0.6. Since these results are in the same range of value as those presented in [2], we can conclude that the VERT-R measure is effective in the summary evaluation.

5. CONCLUSION

In this paper, we have extended ideas from the BLEU and ROUGE algorithms, which are useful in the evaluation of machine translation and text summarization, and proposed the VERT measure for the evaluation of video summaries. VERT-Precision variant has not been found to be effective, while VERT-Recall variant has shown a good correlation with human assessment. With VERT-Recall, several variants have similar performance, so it is hard to choose the best variant. In the future we plan to extend our experiments in size and scope to further identify the capabilities and limitations of the method.

6. Acknowledgement

This research work was partially funded by the ANR Project RPM2.

7. REFERENCES

[1] Papineni, Kishore, 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] C. Lin and E. 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*, Spain, July 25 - 26, 2004.

[4] Paul Over, Alan F. Smeaton, and Philip Kelly, The trecvid 2007 bbc rushes summarization evaluation pilot, *ACM Multimedia*, Germany, September 23–28, 2007.

[5] 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, 2008.

[6] K. Mckeown, R. J.Passonneau and D. K.Elson, Do summaries help? A task-based evaluation of multi-document summarization, *ACM SIGIR conference*, Australia, August 1998.

[7] Yahiaoui, I., Merialdo, B., and Huet, B. 2003. Comparison of multiepisode video summarization algorithms. *EURASIP J. Appl. Signal Process.* 2003 (Jan. 2003), 48-55.

[8] Dumont, E. and Mérialdo, B. 2009. Automatic evaluation method for rushes summary content. In *Proceedings of the 2009 IEEE international Conference on Multimedia and Expo*, New York, NY, USA, June 28 - July 03, 2009.

[9] Yingbo Li and Bernard Merialdo, "Multi-Video Summarization Based on Video-MMR", *International Workshop on Image Analysis for Multimedia Interactive Services*, Desenzano del Garda, Italy, 2010

[10] Dipanjan Das and Andre F,T. Martins, A survey on automatic Text summarization, *Literature survey for the language and statistics II course at CMU*, November 2007.

[11] BT Truong and S. Venkatesh, Video Abstract: A Systematic Review and Classification, *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 3, No. 1, Article 3, 2007