Institut Eurécom
Department of Mobile Communications
2229, route des Crêtes
B.P. 193
06904 Sophia-Antipolis
FRANCE

Research Report RR-10-237

VERT: A METHOD FOR AUTOMATIC EVALUATION OF VIDEO SUMMARIES

April 10th, 2010
Last update April 10th, 2010

Yingbo Li and Bernard Merialdo

Tel : (+33) 4 93 00 81 29
Fax : (+33) 4 93 00 82 00
Email : {Yingbo.Li, Bernard.Merialdo}@eurecom.fr

[1] Institut Eurécom's research is partially supported by its industrial members: BMW Group Research & Technology - BMW Group Company, Bouygues Télécom, Cisco Systems, France Télécom, Hitachi Europe, SFR, Sharp, STMicroelectronics, Swisscom, Thales.

VERT: A METHOD FOR AUTOMATIC EVALUATION OF VIDEO SUMMARIES

Yingbo Li and Bernard Merialdo

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 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, and the best variant is selected through experimentation.

Index Items
Summaries Evaluation, VERT, Rouge, Video Summarization

Contents

List of Figures

List of Tables

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 nowadays. Video summarization [4] [7] [12] 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. Various forms of video summaries are available: static keyframes, video skims and multidimensional browsers. In recent years, multi-video summarization [5] [7] [9] has attracted many researchers. 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 that can be automatically computed, and that shows a strong correlation with human evaluations is therefore of great interest.

Similar situations have already been encountered. In the domain of machine translation [11], BLEU [1] is a popular and 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], especially the basic ROUGE-N and ROUGE-S, counts the n-grams of the candidate summaries co-occurring in the reference summaries to produce an automatically evaluation. In this paper, we propose an algorithm called 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- videos. Red, green and blue being primary colors, ROUGE, VERT and BLEU, their French translations, could become the set of reference evaluation algorithms in their own domains too.

This paper is organized as follows: Section 2 reviews the BLEU/NIST algorithm, the basic theory of ROUGE, and proposes VERT, together with its variants. Section 3 reviews the theory of Video-MMR and experimentally compares several variants of VERT to select the best one. Finally this paper is concluded in Section 4.

2. VIDEO EVALUATION BY RELEVANT THRESHOLD

2.1. BLEU

For the goal of automatically evaluating machine translations, 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. 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.

2.2. ROUGE

Since human evaluation is very time-consuming, a lot of attention in the text summarization area has been devoted to automatic evaluation. The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure proposed by Lin [2] [3] has been proved to be a successful algorithm to complete this task. 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 only those. 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 a set of reference summaries.

ROUGE-N is a recall-related measure, as shown in Eq. 2, while BLEU is a precision-based measure [2]. The number of n-grams in the denominator of Eq. 2 increases if more references are used. When the types of references are changed, the focused aspects of summarization are also changed. And a candidate summary containing different words from more references is favored by ROUGE-N. So it is reasonable for it to prefer a candidate summary with more consensuses with 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, we can limit the maximum skip distance between two in-order words that is allowed to form a skip-bigram.

2.3. VERT (Video Evaluation by Relevant Threshold)

By borrowing ideas from ROUGE and BLEU, we extend these measures to the domain of video summarization. We formalize the process of video summarization as follow:

- We have 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 eventually 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. Depending on the intended format, the video summary may be a concatenated video, or a set of keyframes with specific presentation.

For simplicity, we now assume that keyframes are the basis for selection. Keyframes are 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. By similarity with ROUGE-N, we first propose the VERT-N variant which is defined as:

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

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. 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-N is a recall-related measure too. As ROUGE-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 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, the VERT-N resembles more to the ROUGE-S variant.

2.4. VERT-1 and VERT-2

In this paper, we restrict our study to the cases *n*=1 and *n* =2. We thus define the VERT-1 and VERT-2 measures by Eq. 4 and Eq. 5:

$$\mathit{VERT\text{-}1(C)}= \frac{\sum_{S\in R}\sum_{f\in S} w_C(f)}{\sum_{S\in R}\sum_{f\in S} w_S(f)} \qquad (4)$$

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

In VERT-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. 4 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-2, there are 2 keyframes in each $gram_2$, so Eq. 5 requires a two-dimension computation. We propose two variants for VERT-2:

- VERT-2_S where the weight of a $gram_2$ is the average of the weights of the keyframes:
$$W_S(f,g) = \frac{w_S(f)+w_S(g)}{2}$$
- VERT-2_D, where the weight of a $gram_2$ is the difference between the weights:
$$W_S(f,g) = |w_S(f)-w_S(g)|$$

Obviously, VERT-2_D should only be considered if weights are non-uniform.

The three variants of VERT that we have proposed provide numerical values that are not directly comparable. In order to select which variant seems the most appropriate for consistency with human evaluation, we propose the following approach:

a. We see the problem of video summarization as the selection of a list of k frames (k is fixed) out of a set of K frames. Therefore, there are $K!/(K-k)!$ possible different summaries.
b. Each VERT variant may be normalized so as to be considered as a probability distribution over the set of all possible summaries of length k. The numerical values of different normalized variants may then be compared.

c. A good evaluation measure should assign a high probability to human-generated summaries. The higher the probability (on the average), the better the measure. We can evaluate this average probability by collecting a set of human-generated summaries $H = \{h_1, h_2, \dots, h_r\}$ and compute the normalized value of VERT when h_i is considered as a candidate summary, and $H/\{h_i\}$ is taken as the reference set. Averaging over all h_i provides an estimation of the probability that is assigned to human summaries by the variant. A larger probability will indicate a variant that is more coherent with human selection.

In order to implement this procedure, we first need to compute the normalization factor:

$$NF = \sum_C VERT_R(C) \qquad (6)$$

In Eq. 6, we use the notation VERT_R to explicitly remember the dependence of the VERT measure with respect to the reference set *R*. The summation is taken over all possible summaries *C* of *k* keyframes from the global set of K keyframes. There are very many such summaries, however the computation simplifies nicely:

$$\sum_C VERT_R(C) = \sum_C \frac{\sum_{S\in R}\sum_{g\in S} W_C(g)}{\sum_{S\in R}\sum_{g\in S} W_S(g)} = \frac{\sum_C \sum_{S\in R}\sum_{g\in S} W_C(g)}{\sum_{S\in R}\sum_{g\in S} W_S(g)}$$

$$= \frac{\sum_{S\in R}\sum_{g\in S}\sum_C W_C(g)}{\sum_{S\in R}\sum_{g\in S} W_S(g)}$$

$$= \frac{\left(\sum_C W_C(g)\right)\left(\sum_{S\in R}\sum_{g\in S} 1\right)}{\sum_{S\in R}\sum_{g\in S} W_S(g)}$$

For N=1, $\sum_C W_C(g)$ is simply the product of the sum A of all weights times the number of summaries having g in a given position, namely $(K-1)!/(K-k)!$. Overall, we find that:

$$NF_1 = \frac{A\dfrac{(K-1)!}{(K-k)!}|R|k}{A|R|} = k\frac{(K-1)!}{(K-k)!} \qquad (7)$$

Similar computations apply for VERT-2. Unfortunately, the derivation is longer and we do not have sufficient space to insert it in this paper, but the final result is that the normalization factor is:

$$NF_2 = k(k-1)\frac{(K-2)!}{(K-k)!} \qquad (8)$$

Note that the normalization factor does not depend on the reference set, and is the same for both variants of VERT-2. This is due to the fact that the weights depend on the rank only, and that the contributions of the weights factor in the same manner on both the numerator and the denominator, leaving only a count of combinations left in the final expression.

With these expressions of the normalization factor, we can define a quality value for the VERT variants:

$$Q = \frac{1}{r}\sum_{i=1}^{r}\frac{VERT_{H/h_i}(h_i)}{NF} \qquad (9)$$

where we average the evaluation of each human-selected summary with the other summaries taken as references. The VERT measure with the largest Q value should be the one which achieves the maximum coincidence with the content of human reference summaries.

3. 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 and originating from different sources. “DATI” includes 16 videos, while “YSL” has 14 videos. The “DATI” set contains videos about a French politician woman: most are directly captured from TV news, showing either the person herself, or people commenting her actions. The “YSL” set contains videos related to the death of a famous designer. Some videos represent the burial, some are interviews or comments, some replay older fashions shows. It may happen that some videos are incorrectly classified and unrelated to the topic.

This section is organized as follows: Subsection 3.1 briefly reviews a multi-video summarization algorithm, Video-MMR, whose summary keyframes are used to construct the reference summaries and demonstrate the effect of VERT, and the distances between videos are also defined in this subsection; Subsection 3.2 explains the method of constructing the reference summaries by human selection from the keyframes of subsection 3.1, and two systems of weights are suggested: ranking weights and uniform weights; Subsection 3.3 shows the experimental results of three VERT variants for ranking weights and two VERT variants for uniform weights, decides the final variant, and proves the quality of VERT.

3.1. Video-MMR Algorithm

Video Maximal Marginal Relevance (Video-MMR) [9] is an algorithm to perform video summarization. It builds a summary incrementally by rewarding relevant keyframes and penalizing redundant keyframes. Video-MMR is defined by the recursive formula:

$$S_{k+1} = S_k \cup \underset{f \in V \setminus S_k}{argmax} \begin{pmatrix} \lambda \, Sim_1(f, V \setminus S_k) \\ -(1-\lambda) \underset{g \in S_k}{max} Sim_2(f,g) \end{pmatrix} \qquad (10)$$

where S_k is the current summary, V is the video set, g is a frame inside S_k, and f is a frame inside the set of V except S_k. Sim_1 displays the information between f and the unselected frames $V \setminus S_k$, while Sim_2 shows the information between f and the existing summary S_k.

We also define the distance between two videos by the following formula:

$$d(V_1, V_2) = \frac{1}{n} \sum_{j=1}^{n} \underset{f_j \in V_1, g \in V_2}{min} \left[1 - sim\left(f_j, g\right)\right] \qquad (11)$$

where $sim\left(f_j, g\right)$ is the visual similarity between two frames f_j and g, which are respectively in videos V_1 and V_2.

We will exploit Eq.10 and Eq. 11 in the following subsections.

3.2. Construction of the Reference Summaries

We now detail how we organized the construction of human-selected summaries which will 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.

a. For each video set, we identify 6 representative videos. For this, we use Eq. 11 to compute the mean distance between each video and all the others in the set. We select the 3 videos with the smallest mean, as containing the core of the set, and the 3 videos with the highest mean, as containing the most distinctive information from the set.
b. On these 6 videos, we perform shot boundary detection, and we select one representative keyframe per shot.
c. If a video produces more than 10 keyframes, we select the first 10 keyframes selected 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.
d. 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 textual information that was related to the news items on the original web site.

The reference summaries of video sets "DATI" and "YSL" are shown in Fig. 1 and Fig. 2. Pictures are named from row 1 to 6 from top to down, and column A to J from left to right. The images in the same row originate from the same video. We enrolled 12 users, member of other projects in the laboratory, and their selections are shown in Table.1 and Table. 2, where each row lists the names of the selected pictures for a reference summary.

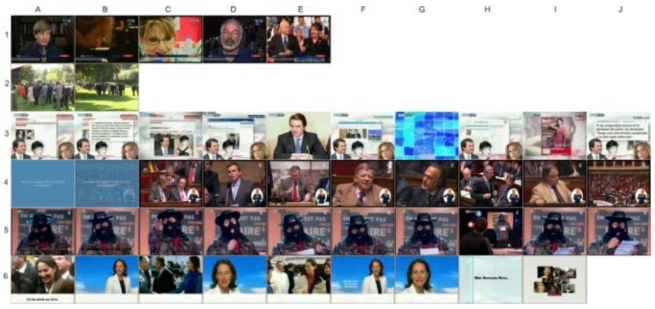

Figure 1. The set of keyframes to construct the reference summaries of "DATI".

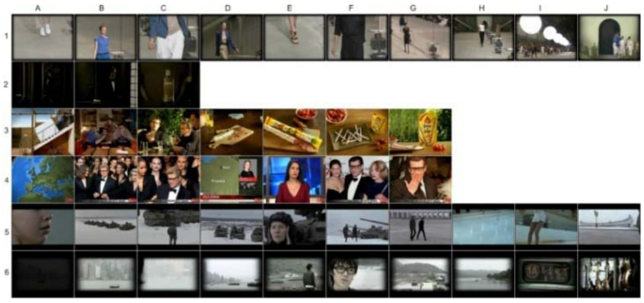

Figure 2. The set of keyframes to construct the reference summaries of "YSL".

Table 1. Reference summaries of "DATI"

5H	2A	2B	3D	3E	4D	4H	6A	6B	1E
1E	1C	2A	3B	3E	4D	4H	5D	5H	6A
3B	1E	5H	2B	1C	3E	4F	5E	6I	3I
1C	1E	2A	3J	4D	4E	5G	5H	6C	6F
6F	5J	4D	4H	3E	1E	2A	6G	3A	5H
3B	2A	5C	4D	1C	3I	5H	4J	6E	1E
3E	5J	1E	2A	4D	6D	4F	3I	5H	6A
2A	3C	4I	5C	1E	6C	3E	6E	3G	5J
1E	2A	3A	3J	4D	4H	5E	5I	6B	6G
4D	4I	1E	2A	6C	4J	3E	5E	4C	5J
1E	2A	3F	4H	5H	6E	6A	1A	3B	4C
2A	3I	3C	3F	3E	1E	1C	5J	5H	6F

Table 2. Reference summaries of "YSL"

1I	1J	4B	4F	6D	6J	5C	3C	3B	2C
1B	1D	2C	3B	3E	4C	4D	4G	5B	5G
4F	3G	1D	3E	5E	6H	4C	2C	1J	5F
1G	1I	2C	3E	4B	4D	5B	5H	6D	6F
6B	5F	4F	3F	2C	1I	6E	5I	4D	3G
4C	1D	2C	1H	5E	6F	4F	3G	3C	6B
2C	3E	3G	1B	4E	4D	5F	6F	4G	5A
2C	3G	3E	4D	5F	5I	6J	6I	4C	1J
1C	1F	2C	3E	4B	4F	5F	5G	6F	6J
1D	4B	1H	5E	6J	1I	2C	3B	3G	6E
1I	2C	3F	4F	5G	6F	3C	4D	5F	6J
4G	4F	1B	1D	1H	1J	2B	6B	6F	6J

3.3. Choice of the Best VERT Measure

In this experiment, we use a set of weights decreasing linearly from 1.0 (for the most important frame) to 0.1 (for the least important). We evaluate the quality value Q as defined in Eq. 9 for the three VERT variants using the human-selected summaries previously described. As a comparison, we also evaluate randomly selected summaries, which should provide a lower quality value.

Fig. 3 and Fig. 4 show 12 values $V_{H/h_i}(h_i)$ in Eq. 9 of VERT-1, VERT-2_S and VERT-2_D for video sets "DATI" and "YSL". The mark "o" presents the human data and the mark "*" is the symbol of random data. Table 3 and Table 4 display the values Q and their variances for the three variants. From these experiments, we observe that the Q value of VERT-2_D is globally larger than that of VERT-2_S and VERT-1, and its variance is less, so VERT-2_D is the best performing measure. For all measures, the value of random selections is less than the value of human selections, which provides some grounds for the soundness of those measures.

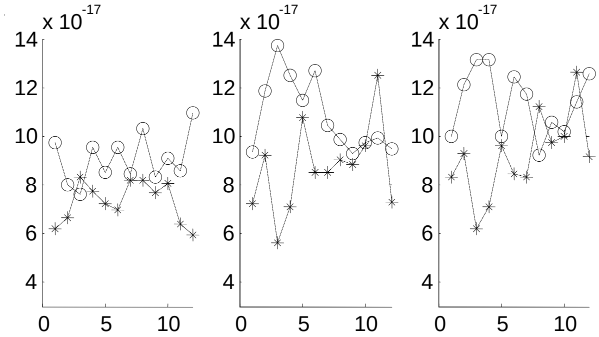

Figure 3. VERT of video set "DATI" for ranking weights.

"o"=human data; "*"=random data.

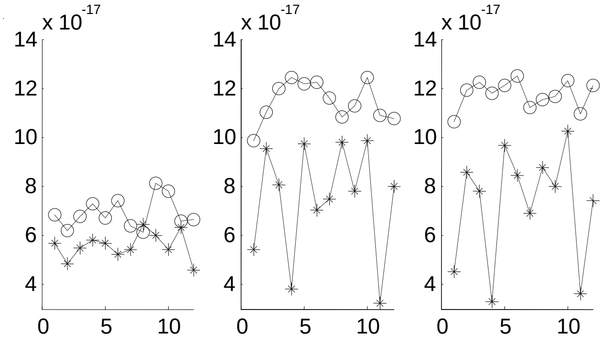

Figure 4. VERT of video set "YSL" for ranking weights.

"o"=human data; "*"=random data.

Table 3. Results Q of VERTs for ranking weights

Q	VERT-1	VERT-2$_\mathrm{S}$	VERT-2$_\mathrm{D}$
DATI(human)	9.0e-017	10.8e-017	11.4e-017
DATI(random)	7.3e-017	8.7e-017	9.2e-017
YSL(human)	6.9e-017	11.5e-017	11.7e-017
YSL(random)	5.6e-017	7.4e-017	7.3e-017

Table 4. Variances belonging to Q for ranking weights

Q	VERT-1	VERT-2$_\mathrm{S}$	VERT-2$_\mathrm{D}$
DATI(human)	1.0e-017	1.4e-017	1.3e-017
DATI(random)	0.8e-017	1.7e-017	1.6e-017
YSL(human)	0.6e-017	0.8e-017	0.5e-017
YSL(random)	0.5e-017	0.2e-017	0.2e-017

We now repeat similar experiments with uniform weights. In this case, VERT-2$_\mathrm{D}$ is not significant, so only VERT-1 and VERT-2 are useful. Fig. 5 and Fig. 6 indicate the different values of $V_{H/h_i}(h_i)$ of VERT-1 and VERT-2 for "DATI" and "YSL". The value Q and the variance belonging to Q of two variants are shown in Table 5 and Table 6. Again, human data provides higher value than random data, and VERT-2 is globally better than VERT-1. Different from uniform weights, the system of ranking weights considers the positions of the keyframes for VERT-1, and the position and relative gap of two keyframes for VERT-2. So the system of ranking system is globally better than the system of uniform weights. However, the system of uniform weights is useful when the ranking weights are hard to decide in some situations.

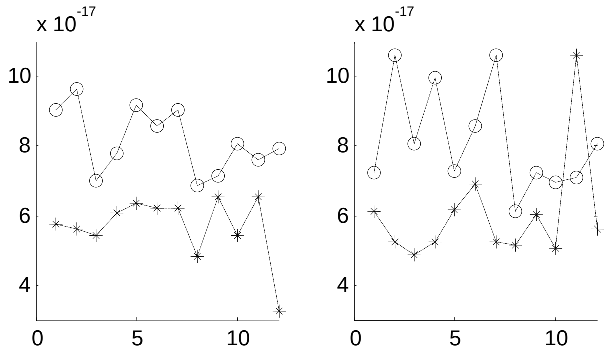

Figure 5. VERT of video set “DATI” for uniform weights.

“o”=human data; “*”=random data.

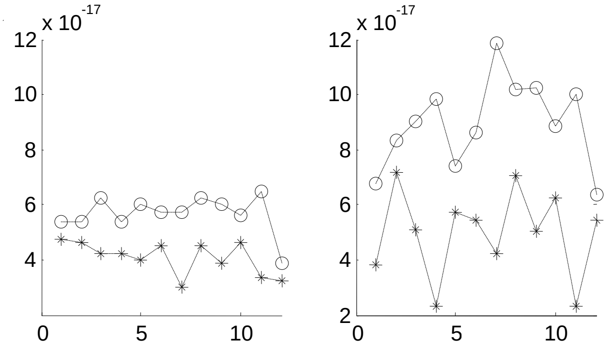

Figure 6. VERT of video set “YSL” for uniform weights.

“o”=human data; “*”=random data.

Table 5. Results $\boldsymbol{Q}$ of VERTs for uniform weights

Q	VERT-1	VERT-2
DATI(human)	8.15e-017	8.16e-017
DATI(random)	5.7e-017	6.1e-017
YSL(human)	5.7e-017	9.0e-017
YSL(random)	4.1e-017	5.0e-017

Table 6. Variances belonging to $\boldsymbol{Q}$ for uniform weights

Q	VERT-1	VERT-2
DATI(human)	0.48e-017	0.79e-017
DATI(random)	0.48e-017	0.82e-017
YSL(human)	0.36e-017	0.82e-017
YSL(random)	0.30e-017	0.85e-017

4. CONCLUSIONS

In this paper, we extend ideas from the BLEU and ROUGE algorithms, which are useful in the evaluation of machine translation and text summarization, and propose the VERT measure for the evaluation of video summaries. VERT is a recall-related measure, as is ROUGE. We introduce three variants of VERT, and explain how they can be compared using a set of human-generated summaries. A quality value for the measures is defined. We describe a set of experiments which compare these three variants, using a reference set of 12 human-selected summaries. The experiments show that the VERT-2_D variant provides the highest quality value when ranking is important, while VERT-2 is better when it is not. Based on the success of BLEU and ROUGE, and the importance of having an automatic evaluation measure for video summaries that is closely related to human evaluation, we believe that VERT has a high potential to participate in a standard for video summarization evaluation. We plan to extend our experiments in size and scope to further identify the capabilities and limitations of the method.

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] 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.

[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, 121-143, February 2008.

[6] Kathleen Mckeown, Rebecca J.Passonneau and David K.Elson, “Do summaries help? A task-based evaluation of multi-document summarization”, ACM SIGIR conference, Melbourne Australia, August 1998.

[7] Hidden for Anonymous reason.

[8] Hidden for Anonymous reason.

[9] Hidden for Anonymous reason.

[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] D. D´echelotte, H. Schwenk, H. Bonneau-Maynard, A. Allauzen and G. Adda, “A state-of-the-art Statistical Machine Translation

System based on Moses”, In Proc. of the tenth MT Summit, pages 451–457, Phuket Thailand, September 2007.

[12] CEES G.M. SNOEK and MARCEL WORRING, “Multimodal Video Indexing: A Review of the State-of-the-art”, Multimedia Tools and Applications, 25, 5–35, 2005.