AUTOMATIC EVALUATION METHOD FOR RUSHES SUMMARY CONTENT

Emilie Dumont and Bernard Mérialdo

Institut Eurécom
Département Communications Multimédia
2229, route des Crêtes -B.P. 193
06904 Sophia-Antipolis cedex - France
{dumont, merialdo}@eurecom.fr

ABSTRACT

During the last years, the developpment of rushes video summarization systems has greatly increased thanks to the international evaluation campaign TRECVID. In this paper, we propose an automation of the manual TRECVID evaluation using machine learning techniques. We train an automatic assessor to perform evaluation on summary content and we show a high correlation between the manual evaluation performed in TRECVID 2008 and our automatic assessor.

***Index Terms*—** Video summarization, TRECVID, evaluation, machine learning

1. INTRODUCTION

Automatic summarization is a useful tool which allows a user to grasp rapidly the essential content of a video without the need to watch the entire document. Automatic video summarization is a challenge since it necessitates decisions about the semantic content and importance of each sequence in a video. This factor complicates the development of automatic video summarization systems and in particular, of evaluation methods. Much of the complexity of summary evaluation arises in the fact that it is difficult to specify what one really needs to measure, without a precise formulation of what the summary is aimed to capture.

For the TRECVID 2008 BBC rushes summarization evaluation campaign [1], authors proposed a manual method to evaluate summaries taking into account conclusions of previous works, like [2, 3, 4, 5]. The quality of each summary is evaluated by objective and subjective metrics: a human judge is given the summary and a chronological list of up 12 topics from a ground truth description of the video content. The assessor views the summary and determines which topics are present. The percentage of topics found by the assessor is the main measure of the summary quality. Other measures are evaluated: did summary contain lots of junk, contain lots of duplicate video and had a pleasant tempo/rhythm? Other indicators are collected in these experiments: duration of the summary, difference between target and actual summary size, total time spent judging the inclusions, total video play time. This approach has the advantage of clearly defining the measures to use for evaluating summaries, and a number of research groups have participated in this task producing summaries suited to this evaluation. The main problem is that this evaluation is currently performed by human judges. This creates fundamental difficulties because evaluation experiments are expensive to reproduce and subject to the variability of human judgment. In particular, this greatly restricts the usage of training methods in the construction of summaries, because they often require a lot of parameter tuning to provide optimal performance.

Our approach is to search for an automation of the evaluation procedure proposed in TRECVID 2008 using the same quality criteria. We decided to focus on the main indicator: the percentage of topics found in the summary IN, because other subjective measures are correlated with it [6]. Previous work already tackled this problem. In [7, 8] authors automated evaluation with a basic and efficient method: a topic is found by the automatic evaluator if a frame sequence of summary overlaps with one of the occurrences of this topic in the original video during one second. The work presented here is an extension of these approaches, in particular [9]. Authors proposed an automation of the manual TRECVID evaluation using machine learning techniques. The main difference is the definition of objects used to predict the topic presence.

2. TRECVID 2008

2.1. BBC rushes video summarization

In the TRECVID 2008 BBC rushes summarization evaluation pilot, the task is to automatically create an MPEG-1 summary clip no longer than 2% of the full video that shows the main objects and events in the video.

2.2. Manual evaluation

2.2.1. Ground truth data

The ground truth is a list of important video segments, each identified by means of a distinctive object or event occurring in the segment with qualifications concerning camera angle, distance, or some other information to make each item description unique. A complete explication can be found in [1]. The ground truth provided by TRECVID is a simple chronological list of topics. This is not sufficient for an automatic evaluation: we augmented the TRECVID ground truth data by manual annotation the test videos to define the precise time segments where each of these topics was present in the video. The average number of ground truth topics for each video is more than 20. In TRECVID, this was considered too large for human evaluators, so that the evaluation was only performed for a random list of 12 topics per video.

2.2.2. Evaluation

Each submitted summary was judged by three different human judges (assessors). An assessor was given the summary and a corresponding list of up 12 topics from the ground truth. The assessor viewed the summary in a $125mm$x$102mm$ mplayer window at 25 frames per second using only play and pause controls and then was asked which of the designated topics appeared in the summary. The percentage of topics found by each assessor determines the fraction of important segments from the full video included. The total score for a summary is the average of the scores given by the three assessors: IN. Figure 1 depicts the video summary evaluation process. The results of the manual evaluation were statistically analyzed in [1], and in conclusion authors found that there was a good agreement between assessor judgments based on the comparison of the topics detected by two assessors in a summary.

Fig. 1. TRECVID 2008 manual evaluation processus

3. AUTOMATIC ASSESSOR

In order to automate the assessment process, we propose to automate the decision on topic detection by creating an automatic assessor so as to be able to automatically predict the topic presence in a summary based on topic, video, and summary features.

3.1. Modelling topic assessment

For our modeling of the automatic assessment, we define a topic instance i as the couple $(\mathbf{x}_i, y_i)$ where:

- $\mathbf{x}_i \in \mathcal{X}$ is a vector containing measurements on the occurrence of the topic,
- $y_i \in \{\text{presence}, \text{absence}\}$ is the result of the decision on the occurrence of the topic, based on the values in $\mathbf{x}_i$.

A topic can have repeated occurrences in a video. We call each of this occurrence a topic sequence. The decision of detecting a topic or not depends on the occurrences of the topic in the original video, and on its occurrences in the proposed summary. Therefore, the vector $\mathbf{x}_i$ which hopefully contains all values necessary to take a decision on a topic presence, contains information coming from the original video and the ground truth, as well as information coming from the proposed summary. In our proposed model, we include the following measurements in the description of a topic instance. The following information is obtained from the augmented ground truth and the original video:

x_1 : Does the topic contain camera motion
x_2 : Does the topic contain an event
x_3 : Number of topic sequences in the video
x_4 : Minimal length of a topic sequence in the video
x_5 : Maximal length of a topic sequence in the video
x_6 : Mean length of a topic sequence in the video
x_7 : Mean activity of topic sequences in the video
x_8 : Mean entropy of topic sequences in the video

The other measurements are obtained automatically from the content of the proposed summary:

x_9 : Number of topic sequences in the summary
x_{10} : Minimal length of a topic sequence in the summary
x_{11} : Maximal length of a topic sequence in the summary
x_{12} : Mean length of a topic sequence in the summary

With this formulation, an automatic assessor will decide on the presence or absence y_i of a topic based on the values of the measurements in $\mathbf{x}_i$.

3.2. Training an automatic assessor

An automatic assessor will define a function $prediction$ that predicts the presence or absence of a topic. If a topic is present, the function returns 1, else the function returns 0. So, once this prediction function is defined by a machine learning, we can compute automatically the IN indicator, the

percentage of topics found in the summary, for a video v by the following formula:

$$IN(v) = \frac{1}{N}\sum_{i=1}^{N} prediction(i)$$

where N is the number of topics in the video. We can also rank summaries according IN.

From the detailed submission results to the TRECVID summarization task, we can create training data in the form of a set of topic instances $(\mathbf{x}_i, y_i)$. This list contains the various decisions y_i made by the assessors on proposed summaries, together with the corresponding measurements x_i on the occurrences of the corresponding topic. Based on this training data, we compare various machine learning techniques to construct an automated assessor, with the objective that it provides decisions that are as close as possible to those of the human assessors.

4. ASSESSOR EVALUATION

In statistics, the Pearson product-moment correlation coefficient r is a common measure of the correlation between two variables. For our problem, Pearson's correlation reflects the degree of linear relationship between manual evaluation and automatic evaluation. It ranges from $+1$ to -1. A correlation of $+1$ means that there is a perfect positive linear relationship between evaluations. A correlation of -1 means that there is a perfect negative linear relationship between evaluations. A correlation of 0 means there is no linear relationship between evaluations.

$$r = \frac{\mathrm{cov}(X,Y)}{\sqrt{(V(X)V(Y))}}$$

where $\mathrm{cov}(X,Y)$ denotes the covariance between X and Y, and $V(X), V(Y)$ respectively the variance of X, and the variance of Y. We evaluate the correlation at 3 different levels :

- Topic: we compute the correlation of the prediction of the topic presence between our automatic assessor and the manual assessors.
- IN: we compute the correlation of the IN evaluation between our automatic evaluation and the manual evaluation.
- Ranking: we compute the correlation of ranking of summaries according IN between our automatic evaluation and the manual evaluation.

5. EXPERIMENTAL RESULTS

We experimented our approach on 8 videos proposed by BBC Rushes Task 2008 in TRECVID. It consists of unedited video footage, shot mainly for five series of BBC drama programs and was provided to TRECVID for research purposes by BBC archive. The training instances are obtained from results of TRECVID 2008 evaluation for 10 summarization systems kindly sent to us by various participants.

In order to have the training data and the training test completely independent, we choose to use a leave-one-out technique: for each video and for each system, we train the model on 7 videos x 9 systems and we test the model on the last video and system. Figure 2 shows the Pearson correlation between manual assessor and our automatic assessor at 3 levels (topic, IN and ranking) for different machine learning methods using the software WEKA.

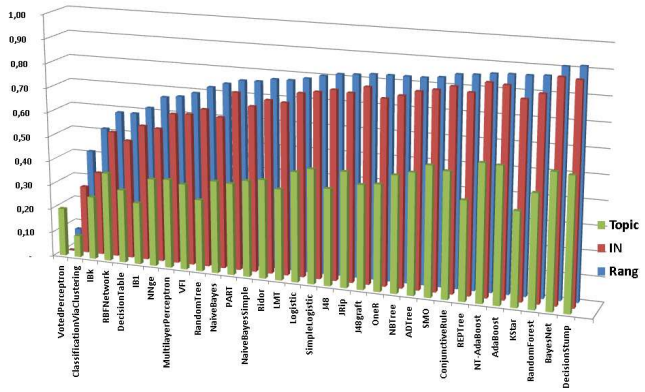


Fig. 2. Pearson correlation between manual and automatic assessors.

The difference between classifiers is large, so it is important to choose a classifier according to our problem: the decision stumps. A decision stump is a weak machine learning model consisting of a decision tree with only a single depth. To perform evaluation, we train several decision stumps based on the leave-one-out method. We obtain only 4 different decision stumps, figure 3 shows models we found.

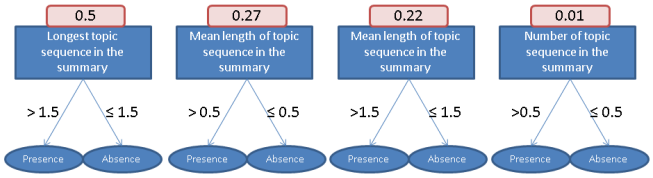


Fig. 3. Decision stumps found.
At the top, the ratio of occurence of the decision stump.

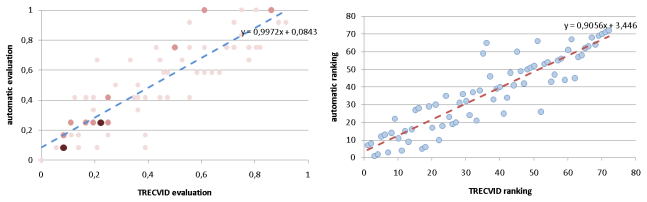
The table 1 shows the manual prediction of presence topic according to the automatic prediction, the Pearson coefficient at this level is 0.54, so the correlation is moderate.

Figure 4 shows the manual evaluation of IN according to the automatic evaluation and the manual evaluation of rank-

Classified as	Absence	Presence
Absence	390	116
Presence	45	313

Table 1. Confusion matrix of the topic presence prediction.

ing according to the automatic evaluation. At these levels, the Pearson correlation is 0.88 and 0.91, so the correlation is high.



(a) IN level (b) Ranking level

Fig. 4. Pearson correlation between manual assesor and our automatic assessor

In reality, manual assessors have not the same judgment, because of subjective interpretation of topic occurrence. We would like a classifier that shows a close agreement with manual evaluation, if possible closest between two human assessors. We evaluate quality of our automatic assessor in comparison to an human assessor. We use all manual evaluations done by TRECVID. For each pair of assessors, we compute the correlation between their evaluations at the 3 levels. We average coefficients for each assessor, table 2 shows the results.

Assessor	Topic	IN	Ranking
Assessor 1	0.755961	0.878687	0.914713
Assessor 2	0.789278	0.875702	0.917511
Assessor 3	0.770808	0.870860	0.911205
Assessor 4	0.775011	0.860053	0.895711
Assessor 5	0.790169	0.865818	0.897900
Assessor 6	0.750509	0.805306	0.833046
Assessor 7	0.715957	0.781069	0.824572
Assessor 8	0.702580	0.804130	0.811149
Assessor 9	0.726755	0.855810	0.892882
Assessor 10	0.790546	0.901866	0.926379
DecisionStump	0.535261	0.875906	0.913306

Table 2. Assessor correlation.

Experiments show that the method proposed to automatically evaluate summary video is almost as good as evaluation performed by human assessors: the correlation between these evaluations is high as the correlation between manual evaluations. But at the topic level, the automatic assessor has only a moderate correlation with the manual evaluation. So, our experiments demonstrate that our automatic evaluation technique is suitable for comparing and evaluating summaries using IN indicator.

6. CONCLUSION

We have proposed an approach to automate the summary evaluation by training a decision stump in order to remove the human interaction that was required in the TRECVID evaluation campaign. Through experiments, we showed a high correlation between the manual evaluation proposed by TRECVID2008 and our automatic evaluation. In further work, it would be interesting to generalize our approach on a larger data set including more videos and more summarization systems to improve the prediction quality.

7. REFERENCES

[1] Paul Over, Alan F. Smeaton, and George Awad, "The TRECVID 2008 BBC rushes summarization evaluation," in *Proceedings of the TRECVID Workshop on Video Summarization (TVS'08)*, Vancouver, BC, Canada, October 2008.

[2] A.M. Ferman, A.M.; Tekalp, "Two-stage hierarchical video summary extraction to match low-level user browsing preferences," *Multimedia, IEEE Transactions on*, vol. 5, no. 2, pp. 244–256, June 2003.

[3] A. Ekin, A.M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *Image Processing, IEEE Transactions on*, vol. 12, no. 7, pp. 796–807, July 2003.

[4] Cuneyt M. Taskiran, "Evaluation of automatic video summarization systems," in *Multimedia Content Analysis, Management, and Retrieval 2006*, Edward Y. Chang, Alan Hanjalic, and Nicu Sebe, Eds. 2006, SPIE.

[5] Ba Tu Truong and Svetha Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, no. 1, 2007.

[6] Marcin Detyniecki and Christophe Marsala, "Adaptive acceleration and shot stacking for video rushes summarization," in *Proceedings of the TRECVID Workshop on Video Summarization (TVS'08)*, Vancouver, BC, Canada, October 2008.

[7] Emilie Dumont and Bernard Mérialdo, "Split-screen dynamically accelerated video summaries," in *MM 2007, 15th international ACM conference on multimedia, September 24-29, 2007, Augsburg, Germany*, Sep 2007.

[8] Alexander G. Hauptmann, Michael G. Christel, Wei-Hao Lin, Bryan Maher, Jun Yang, Robert V. Baron, and Guang Xiang, "Clever clustering vs. simple speed-up for summarizing rushes," in *TVS '07: Proceedings of the international workshop on TRECVID video summarization*, New York, NY, USA, 2007, pp. 20–24, ACM.

[9] Dumont Emilie and Mérialdo Bernard, "Automatic evaluation method for rushes summarization: experimentation and analysis," in *CBMI 2008, 6th International Workshop on Content-Based Multimedia Indexing, June 18-20, 2008, London, UK*, 2008.