AUTOMATIC EVALUATION METHOD FOR RUSHES SUMMARIZATION: EXPERIMENTATION AND ANALYSIS

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

Evaluation remains an important difficulty in the development of video summarization systems. Rigorous evaluation of video summaries generated by automatic systems is a complicated process because the ground truth is often difficult to define, and even when it exists, it is difficult to match with the obtain results.
The TRECVID BBC evaluation campaign has recently introduced a rushes summarization task and has defined a manual evaluation methodology. In this paper, we explore the use of machine learning techniques to automate this evaluation. We present our approach and describe the current results, in comparison with manual evaluations performed in the 2007 campaign.

1. INTRODUCTION

Recent technological improvements have greatly increased computing power, electronic storage capacity and transmission bandwidth. Multimedia information and particularly digital video is becoming more and more common and very important for education, entertainment and many other applications. This large amount of multimedia data has fueled efforts to provide and develop techniques for efficiently processing and manipulating this type of data. In this paper, we focus on video summarization, and in particular on the difficult problem of video summarization evaluation.

Automatic summarization is a useful tool which allows a user to grasp rapidly the essential content of a video, without the need for watching the entire document. Automatic video summarization is a challenge since required to make decisions about the semantic content and importance of each sequences 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.

In the following section, we explain our motivation and present a brief review of recent approaches to summary evaluation. Section 3 details our method to automatically evaluate summaries. And finally, we propose an experimentation in section 4 and an analysis of this method in section 5.

2. MOTIVATION

The ever-growing availability of videos, creates a strong requirement for efficient tools to manipulate and present this data in an effective manner. Automatic summarization is one of those tools. The idea is to automatically and without human interaction create a short version which contains as much information as possible as in the original video. The key issue here is to define what should be kept in the summary and how this relevant information can be automatically extracted. A number of approaches have been proposed to define and identify what is the most important content in a video [1], [2], [3].

A major problem to develop summaries is the fact that evaluation is difficult, in the sense that it is hard to judge the quality of a summary, or, when a performance measure is available, it is hard to understand what its interpretation is. So, in the field of automatic summarization, most papers suggest they own evaluation technique, chosen appropriately for their own task. This makes the comparison of different systems difficult, if not impossible, and creates an urgent need for a commonly accepted evaluation methodology.

In the TRECVID 2007 BBC rushes summarization evaluation pilot [4], authors propose a manual method to evaluate summaries taking into account conclusions of previous works, like [5], [6], [2], [7]. 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 indicators are collected in these experiments: ease of finding desired content as judged by assessor, amount of near redundancy as judged by assessor, assessor time taken to determine presence/absence of desired segments, size of summary and elapsed time for summary creation. 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. So, our approach is to search for an automation of the evaluation procedure proposed in TRECVID, using the same quality criteria. Previous work already tackled this problem: in [8], [9], 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.

3. AUTOMATING THE EVALUATION

We decided to focus on the performance main indicator IN: the percentage of topics found in the summary. Other indicators used in TRECVID are related to usability, which provides a different aspect of evaluation.

3.1. Ground truth data

The initial conception of ground truth was a list of objects and events, but the application of this view to even a small sample of the data quickly made it clear that there would be too many such items. As a result the working notion of ground truth was changed to be 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 find in [4]. The ground truth provided by TRECVID is a simple chronological list. This is not sufficient for an automatic evaluation, so, for our purpose, we augmented the TRECVID ground truth data and manually annotated the test videos to define the time segments where each of these topics was present in the video.

The topics describe portions of the video showing people, objects, events, locations, ... and combinations of the former, sometimes combined with camera motion information. An example of augmented ground truth list, containing the topic descriptions and the time segment boundaries, is shown in figure 1. 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. We followed the same process in our experiments, and for each video, we produced a sublist of 12 topics chosen at random.

```
* 1 blonde haired man and woman with red hair
763 780

* 1 blonde haired man walks out of the room
780 855

* 1 woman in black suit at coffee pot,
  blonde haired man in background
1032 1395

* 1 camera pans by woman in black suit,
  to blonde haired man and woman with red hair
1395 1600

* 1 blonde haired man hands tapes to woman in red shirt
2140 2263

# 4 woman in red shirt watches tv
2346 2420
2512 2666
2725 3220
3352 3528

* 2 blonde haired man walks to woman in red shirt watching tv
2420 2511
3220 3352

* 3 woman with blue shirt put videos in a box
3530 3774
3849 3975
4452 4574

* 1 man with tie enters room with woman with blue shirt
3975 4050

* 1 man with tie talks to woman with blue shirt
4051 4300

* 1 man with tie leaves room,
  leaving woman with blue shirt and man with a blonde hair
4301 4355

* 1 man with blonde hair talks to woman with blue shirt
4356 4400

* 1 man with tie leaves room, leaving woman with blue shirt
4401 4450
```

Fig. 1. Augmented Ground truth of video MRS043400

3.2. Manual evaluation in TRECVID 2007

Each submitted summary for each of the 42 test videos 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*102 mm mplayer window at 25 frames per second using only play and pause controls and then determined which of the designated topics appeared in the summary. The percentage of topics found by 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. The results of the manual evaluation were statistically analyzed in [4], and in conclusion authors found that there was a strong agreement between assessor judgments, based on the comparison of the topics detected by two assessors in a summary.

3.3. Automatic assessor

In order to automate the assessment process, we propose to automate the decision on topic detection, so as to be able to automatically compute IN, the percentage of topics found, using machine learning techniques.

3.3.1. 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 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, 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:

- Video id
- Number of sequences (of this topic) in this video
- Minimal length of a sequence in this video
- Maximal length of a sequence in this video
- Mean length of a sequence in this video
- Mean activity of sequences in this video
- Mean entropy of sequences in this video

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

- Number of sequences (of this topic) in the summary
- Minimal length of a sequence in the summary
- Maximal length of a sequence in the summary
- Mean length of a 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 $\mathbf{x}_i$.

3.3.2. 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, 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.

From the detailed results of our submission 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 our 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. The supervised methods that we used are listed in table [**?**], we used the software library WEKA [10] for our tests.

3.4. Evaluation of automatic assessment

The goal of our experiment is to automatically compute IN. So, we would ideally expect a high correlation, and agreement between manual and automatic evaluation, with a variability between the automatic assessor and human assessors that equals the variability between human assessors.

Alternating Decision Tree [11]	ADTree
Bayes Network learning	BayesNet
simple meta-Classifier Via clustering	CVClustering
single Conjunctive Rule	ConjunctiveRule
Decision Stump	DecisionStump
simple Decision Table majority [12]	DecisionTable
HyperPipe classifier	HyperPipes
K-nearest neighbours [13]	IBk
C4.5 Decision Tree [14]	J48
repeated incremental pruning [15]	JRip
instance-based classifier [16]	KStar
Lazy Bayesian Rules [17]	LBR
Logistic Model Trees	LMT
multinomial Logistic regression [18]	Logistic
Multilayer Perceptron	MultilayerPerceptron
Naive Bayes [19]	NaiveBayes
Simple Naive Bayes [20]	NaiveBayesSimple
decision Tree with Naive Bayes classifiers at the leaves [21]	NBTree
Nearest-neighbor-like algorithm using non-nested generalized exemplars [22]	NNge
uses the minimum-error attribute for prediction [23]	OneR
PART decision list [24]	PART
Forest of Random trees [25]	RandomForest
a tree that considers K randomly chosen attributes at each node	RandomTree
Normalized Gaussian Radial Basis function Network	RBFNetwork
Fast Decision Tree	REPTree
RIpple-DOwn Rule learner. [26]	Ridor
Logistic Model Trees	SimpleLogistic
Sequential Minimal Optimization	SMO
Voting Feature Intervals	VFI
Voted Perceptron	VotedPerceptron
0-R classifier	ZeroR

Table 1. Supervised learning methods

3.4.1. Correlation

In statistics, the Pearson product-moment correlation coefficient r is a common measure of the correlation between two variables. Pearson's correlation reflects the degree of linear relationship between two variables X and Y with n datas. It ranges from $+1$ to -1. A correlation of $+1$ means that there is a perfect positive linear relationship between variables. A correlation of -1 means that there is a perfect negative linear relationship between variables. A correlation of 0 means there is no linear relationship between the two variables.

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

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.

3.4.2. Variability

The mean squared error MSE of an estimator is one of many ways to quantify the amount by which an estimator differs from another. MSE measures the average of the square of the "error" between two evaluations. The error is the amount by which the first evaluation differs from the second to be estimated, so a MSE of zero, meaning that the two evaluations compute IN with perfect accuracy.

$$MSE = \frac{1}{V}\sum_{v=1}^{N}(IN_a(v) - IN_b(v))^2$$

where V denotes the number of videos.

3.4.3. Agreement

Cohen's kappa coefficient is a statistical measure of inter-rater reliability. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. Cohen's kappa measures the agreement between two evaluation who each classify N topics into absence or presence, mutually exclusive categories. The equation for κ is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

where $Pr(a)$ is the relative observed agreement among evaluations, and $Pr(e)$ is the probability that agreement is due to chance. If the evaluations are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa = 0$.

4. EXPERIMENTS

4.1. Video data

We experimented our approach on 7 videos proposed by BBC Rushes Task 2007 in TRECVID, see table 2. 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 the detailed results of TRECVID 2007 evaluation, for 3 summarization systems:

- Baseline1 system is an uniform baseline of 4% summaries in which one second was selected for every 25 seconds of original video.
- Baseline2 used a shot boundary detection, for each shot a keyframe is extracted. All keyframes were used in a K-means clustering, which the number of clusters set to the number of seconds in the 4% summary. From each cluster, the single shot closest

to the centroid was selected, and one second from the middle of this shot is used for inclusion in the summary.

- In Eurecom system [8], video is segmented into shots, and the most important and non-redundant shots were selected for inclusion in the summary. During presentation in the summary, shots were dynamically accelerated according to the motion activity, and presented using a split-screen display.

	MRS035132	MRS150148	MRS157475	MS237650	MRS043400	MRS157443	MS212920
Video	326	1663	1556	198	782	2181	833
Baseline1	13	68	64	7	32	90	34
Baseline2	12	67	63	7	31	89	33
Eurecom	3	29	51	7	27	49	11

Table 2. Duration in seconds of summaries

4.2. Classifiers

The first step is to identify the best classifiers to predict the percentage of topics found. We consider that the manual evaluation is close to the perfect evaluation, so we search for classifiers with the best trade-off between:

- a great correlation with manual evaluation
- a great agreement with manual evaluation
- a weak variability with manual evaluation

For each classifier, for each video, and for each system, we estimate the IN indicator by training the classifier on 6 videos and testing on the last one. Figure 2 shows correlation according to agreement for all classifiers.

The difference between classifiers is great, so it is important to choose a classifier according to our problem. We want to estimate the percentage of topics found in a summary, figure 3 shows a zoom of curve 2. We see that a classifier like ADTree, BayesNet, Decision Stump is a good classifier for this problem. ADTree is the classifier with the weakest variability, Decision Stump has the greatest agreement, and BayesNet the greatest correlation with manual evaluation.

4.3. Manual and automatic evaluation

In reality, assessors do not have 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.

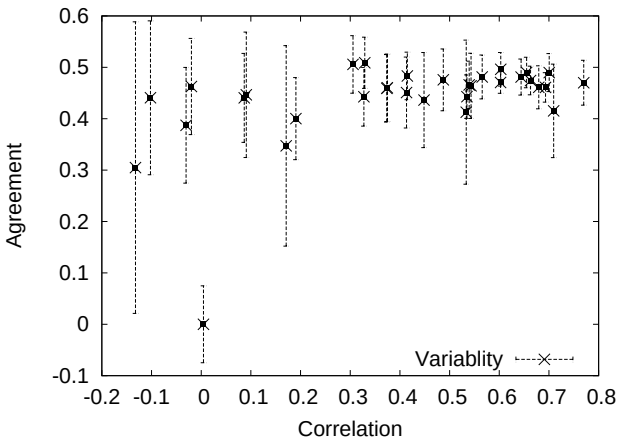


Fig. 2. Correlation according to agreement for all classifiers computed on 21 estimations of IN (7 videos $*$ 3 systems).

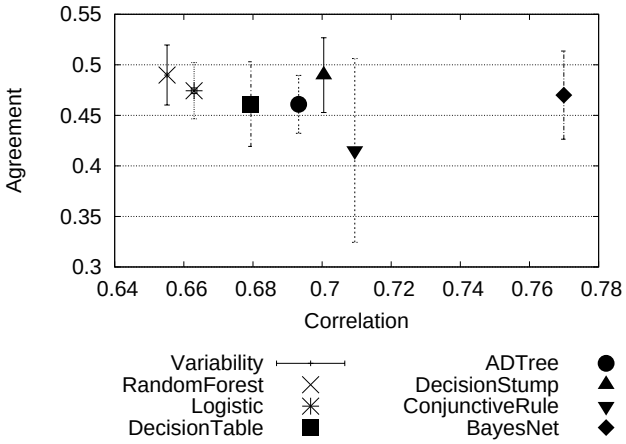
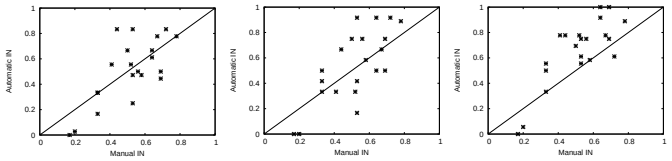


Fig. 3. Correlation according to agreement for classifiers

Figure 4 shows the manual evaluation according to the automatic evaluation for a classifier. The Pearson's coefficient is equals to 0.693, 0.7, 0.77 respectively, for ADTree, DecisionStump and BayesNet, this indicates a marked degree of correlation. Tha Kappa's coefficient is equals to 0.461, 0.49, 0.47, this indicates a moderate agreement.

Now, we evaluate quality of our automatic assessor in comparison to an human assessor. We use all manual evaluations done by TREVCVID, eg for 42 videos and for 24 systems (1005 summaries), it represents 3645 values of IN. For each pair of assessors, we compute the correlation, agreement and variability between their evaluations of IN. We average coefficients for each assessor, table 3 shows the results.

It is clear that human assessors have a good agreement to predict the presence of a topic. In comparison, automatic assessors only have a moderate agreement. But, at the same time, results show that the correlation and variability



(a) ADTree (b) Decision Stump (c) BayesNet

Fig. 4. Manual evaluation compared with automatic evaluation

Assessor	Pearson	Kappa	MSE
Assessor 1	0.738	0.728	0.025
Assessor 2	0.532	0.668	0.046
Assessor 3	0.587	0.640	0.035
Assessor 4	0.607	0.700	0.038
Assessor 5	0.652	0.638	0.035
Assessor 6	0.658	0.683	0.032
Assessor 7	0.637	0.722	0.031
ADTree	0.69	0.461	0.028
BayesNet	0.77	0.49	0.043
DecisionStump	0.7	0.47	0.037

Table 3. Pearson's coefficient, Kappa's coefficient and MSE for assessors

between human assessors and automatic assessors have the same order of magnitude. This allows us to say that the automatic prediction that we propose can successfully be used to compare systems.

5. ANALYSIS

In this section, we further analyze the classifiers which provide the best results: ADTree, DecisionStump and BayesNet. Tables 4 and 5 shows comparisons between these classifiers.

	ADTree	DStump	BayesNet
Correctly Classified Instances %	76.19	70.83	71.88
Incorrectly Classified Instances %	23.8	29.17	28.12
Kappa statistic	0.53	0.43	0.44
Mean absolute error	0.36	0.4	03
Root mean squared error	0.4	0.45	0.45
Relative absolute error %	74	80.19	60.03
Root relative squared error %	81.38	89.55	89.88

Table 4. Error on training data

5.1. Decision Stump

A Decision Stump is a weak machine learning model consisting of a Decision Tree with only a single depth. Figure 5

	ADTree		DStump		BayesNet	
Class	0	1	0	1	0	1
TP Rate	0.81	0.72	0.84	0.60	0.71	0.73
FP Rate	0.28	0.19	0.4	0.16	0.27	0.29
Precision	0.70	0.83	0.64	0.82	0.70	0.73
Recall	0.81	0.72	0.84	0.60	0.71	73
F-Measure	0.75	0.77	0.72	0.69	0.71	0.73
ROC Area	0.85	0.85	0.72	0.72	0.81	0.81

Table 5. Detailed Accuracy by class (0 = absence, 1 = presence)

shows the model trained on 7 videos, for 3 systems and for different assessors, corresponding to 843 topic instances.

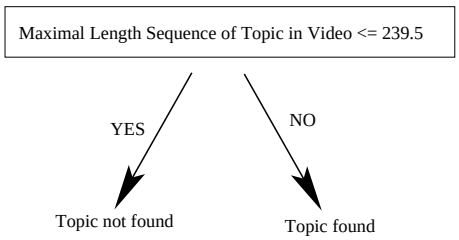


Fig. 5. Decision Stump

We see that the prediction is based on a video attribute. This is caused by the Baseline1 system, which selects one second every 25 seconds from the initial video. As this is not typical of an elaborate summarization mechanism, we restricted the analysis to the only Eurecom system. Decision Stump trained on only this system is presented in 6.

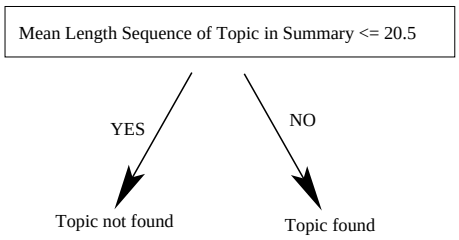


Fig. 6. Decision Stump trained on Eurecom system

This Decision Tree is close to the automatic evaluation methods that had been proposed previously in [8] and [9]: the first considers that the topic is present if one second of the topic was found in the summary, the second that the topic is found if 25 frames are in the summary. The Decision Tree method therefore proposes a very similar rule, with the specific aspect that the threshold has been automatically estimated.

5.2. Alternating Decision Tree

The Alternating Decision Tree combines Decision Trees and boosting. Figure 7 shows the tree trained of all 7 videos for Eurecom system.

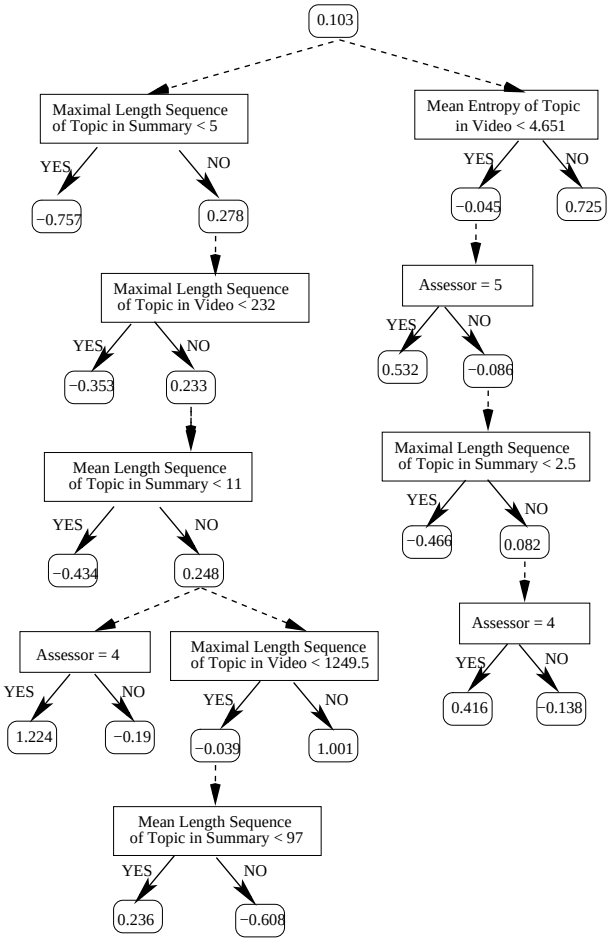


Fig. 7. Alternating Decision Tree trained on Eurecom system. Negative numbers correspond to predicted absence.

This method takes into account many more attributes, and it also assign weights to these attributes. This allows a more detailed analysis of the importance of attributes. It is clear that certain attributes are very important, such as maximum and average length of the sequences in the video and in the summary. The identity of the assessor also can influence the prediction. But on the other hand, it is clear that an attribute such as average activity has no impact on the prediction.

5.3. Bayes Network

A Bayesian network is a probabilistic graphical model that represents a set of variables and their probabilistic independencies. This method uses also many attributes, see table 6. The prediction depends on the video and the assessor, it takes also in account attributes depending of the summary: number of sequences and mean length of these sequences. It also uses the value of maximum length of the sequences. Certain attributes are not used for the prediction, such as the mean activity of the topic in the video, or the mean entropy.

Probability Decision Table for	Value	Presence	Absence
Class		0.476	0.524
Video	MRS157443	0.108	0.136
	MRS035132	0.087	0.041
	MRS150148	0.17	0.076
	MRS157475	0.128	0.117
	MS237650	0.094	0.149
	MRS043400	0.015	0.244
	MS212920	0.274	0.212
Assessor	1	0.208	0.202
	2	0.095	0.074
	3	0.299	0.234
	4	0.187	0.298
	5	0.06	0.106
	6	0.067	0.022
	7	0.074	0.054
Max Length Sequence of Topic in Video	< 239.5	0.554	0.446
	≥ 239.5	0.273	0.727
Number of Sequence of Topic in Summary	< 1	0.362	0.638
	≥ 1	0.034	0.966
Number of Sequence of Topic in Summary	< 2.5	0.409	0.034
	$\geq 2.5 \wedge < 131.5$	0.587	0.825
	≥ 131.5	0.825	0.141
Mean Length Seq of Topic in Summary	< 3	0.427	0.042
	≥ 3	0.573	0.958
Others	All	1	1

Table 6. Probality decision table of Bayes Network trained on Eurecom system.

6. DISCUSSIONS

It is clear that usage of classifiers for the prediction of the presence of a topic in a summary video is efficient. The quality of the prediction for the topic is not very precise, but on the other hand the automatic calculation of the percentage of topics found in a summary has a good correlation with the manual evaluation. This method is therefore usable to compare the relative quality of summaries.

We also can say that the best classifier for this evaluation was found to be the Bayes network. But Decision Stump is also a good classifier for this problem, as well as Alternating Decision Tree.

This analysis allows us to show that some attributes are more important for the prediction of the presence of a topic. We can say that the length of the sequences of the topic in the summary has an essential role in the evaluation, while the

features of the sequences in the video generally do not bring much information.

7. CONCLUSION

We have proposed an approach to automate the evaluation of summaries generated by automatic video summarization systems, in order to remove the human interaction that was required in the TRECVID evaluation campaign. Through experiments, we showed a correlation between the manual evaluations proposed by TRECVID2007 and our automatic evaluation. If it is difficult to build an automatic classifier that provides the same results as a human assessor, a strong result of our experiments is that automatic classifiers are able to accurately compare summarization systems, at least from a relative performance perspective.
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 quality of the prediction, evaluate the effect of new attributes such as for example, a classification of the topics into action, camera motion, dialogs, and validate the effectiveness of automatic assessors when compared with human assessors. This is of important value to facilitate the development of more accurate summarization systems.

8. REFERENCES

[1] H. Jing, R. Barzilay, K. McKeown, and M. Elhadad, "Summarization evaluation methods experiments and analysis," .

[2] Cuneyt M. Taskiran, "Evaluation of automatic video summarization systems," 2006, SPIE.

[3] M.G. Christel, "Evaluation and user studies with respect to video summarization and browsing," .

[4] Paul Over, Alan F. Smeaton, and Philip Kelly, "The TRECVID 2007 BBC rushes summarization evaluation pilot," in *Proceedings of the TRECVID Workshop on Video Summarization (TVS'07)*, New York, NY, September 2007, pp. 1–15, ACM Press.

[5] 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.

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

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

[8] 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.

[9] 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.

[10] http://www.cs.waikato.ac.nz/ ml/weka/, ," .

[11] Y. Freund and L. Mason, "The alternating decision tree learning algorithm," in *Proceeding of the Sixteenth International Conference on Machine Learning*, Bled, Slovenia, 1999, pp. 124–133.

[12] Ron Kohavi, "The power of decision tables," in *8th European Conference on Machine Learning*. 1995, pp. 174–189, Springer.

[13] D. Aha and D. Kibler, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.

[14] Ross Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[15] William W. Cohen, "Fast effective rule induction," in *Twelfth International Conference on Machine Learning*. 1995, pp. 115–123, Morgan Kaufmann.

[16] John G. Cleary and Leonard E. Trigg, "K*: An instance-based learner using an entropic distance measure," in *12th International Conference on Machine Learning*, 1995, pp. 108–114.

[17] Zijian Zheng and G. Webb, "Lazy learning of bayesian rules," *Machine Learning*, vol. 4, no. 1, pp. 53–84, 2000.

[18] S. le Cessie and J.C. van Houwelingen, "Ridge estimators in logistic regression," *Applied Statistics*, vol. 41, no. 1, pp. 191–201, 1992.

[19] George H. John and Pat Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, 1995, pp. 338–345, Morgan Kaufmann.

[20] Richard Duda and Peter Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[21] Ron Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid," in *Second International Conference on Knoledge Discovery and Data Mining*, 1996, pp. 202–207.

[22] Sylvain Roy, "Nearest neighbor with generalization," 2002.

[23] R.C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63–91, 1993.

[24] Eibe Frank and Ian H. Witten, "Generating accurate rule sets without global optimization," in *Fifteenth International Conference on Machine Learning*, J. Shavlik, Ed. 1998, pp. 144–151, Morgan Kaufmann.

[25] Leo Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[26] Brian R. Gaines and Paul Compton, "Induction of ripple-down rules applied to modeling large databases," *J. Intell. Inf. Syst.*, vol. 5, no. 3, pp. 211–228, 1995.