Towards a Standard Protocol for the Evaluation of Temporal Video Segmentation Algorithms

Rosa Ruiloba[1], Philippe Joly[1], Stéphane Marchand-Maillet[2]
and Georges Quénot[3]

[1] Multimedia Team,ASIM-LIP6-Universite Pierre et Marie Curie,
4 Place Jussieu, Paris, FRANCE
{Rosa.Ruiloba, Philippe.Joly}@lip6.fr
[2] Eurécom, Sophia-Antipolis, Nice, FRANCE
Stephane.Marchand@eurecom.fr
[3] CLIPS-IMAG, BP53, 38041 Grenoble Cedex 9, FRANCE
Georges.Quenot@imag.fr

Abstract. This paper proposes a general framework to evaluate and compare several temporal video segmentation algorithms. The problems that must be solved to confront different methods summarise as gathering a common content set to test the methods, building a reference segmentation, establishing the rules to match the results with the reference, and providing a quality measure. Some solutions to these problems are given in this study and are applied for evaluating different methods developed in various contexts. The paper concludes by presenting results obtained on practical tests.

1 Introduction

This paper addresses the problem of evaluating temporal video segmentation algorithms and systems. Among several others, fields where temporal video segmentation techniques find applications are video document indexation and retrieval, information and emission-type filtering and video document browsing. Temporal segmentation must be distinguished from spatial segmentation (object-based segmentation) and spatio-temporal segmentation (object tracking), which are not considered here. This study focuses solely on techniques performing the segmentation of the image track of a video document. This segmentation process mostly consists of detecting "transition effects" between "homogeneous segments", the definition of which being rather application-dependent.

We identify the four following major problems that must be accounted for in the evaluation of temporal video segmentation systems:

- Selection and gathering of a corpus (test database),
- Definition of a reference segmentation for this corpus (ground truth),
- Defining one or more "quality measures" criteria (measure),
- Comparing the automated segmentation and the reference segmentation (evaluation).

Temporal video segmentation techniques can be used in a wide range of applications, each of which inducing various requirements. Examples of such constraints can be described in terms of the type of the transition effects to be recognised, the accuracy of their detection and location, and also in terms of computational times induced. All these constraints have an impact on the above mentioned problems, namely the selection of a corpus, the definition of the effects to be indexed, the definition of the rules for the relevant errors, and the selection of an appropriate quality criterion. There is therefore a strong need (currently not satisfied) for the definition of a global evaluation protocol allowing one to test consistently various algorithms and systems for different targeted applications. Moreover, the experience of developing such evaluation protocols within the field of speech recognition [4] suggests that such a development will raise a set of questions on the problem of temporal video segmentation itself.

In this paper, we first review in detail each of the above issues and discuss their possible solutions. While doing so, we propose a formal context in which temporal video segmentation algorithms will be evaluated. An example of such a validation process is given in the last section.

2 Selection and Gathering of a Corpus

The parameters which characterise a given test corpus include its size, its homogeneity, its complexity and the density of the various types of transition effects it contains. The selection of a corpus is rather independent from the evaluation protocol since it mostly depends on application in question. However, the protocol must take into account all the characteristics of the corpus for the evaluation of the difficulty of performing its segmentation.

The results presented in this paper were obtained using the corpus AIM developed within the French inter-laboratory research group ISIS. This corpus contains 8 video documents for a total duration of 172 minutes. It includes several TV news, various advertising, and a TV series. Other corpora exist such as the MPEG-7 content set [9], which may also be considered in an evaluation process.

3 Definition of a Reference Segmentation

Once the video sequence is segmented using the automated technique, the results must be evaluated. One possibility is to use a human "evaluator" who checks whether the gathered keyframes (the last frame of a shot and the first frame of the next shot, for example) effectively delimit a transition effect or belong to the same shot. The complete video sequence is then used to check the presence of detected effects and to mark the possibly missed transition effects.

The fact that the reference segmentation is performed by a human candidate implies that subjectivity cannot be avoided within this reference indexation. However, it is crucial that subjectivity is fully excluded from the evaluation process. In order to reduce the human error in the manual validation, several evaluators must work concurrently on the same segmentation task. This procedure is very expensive but shows to make the validation very reliable.

The result may also be influenced by the segmentation method and the application domain and may consider only a few types of effects. Therefore, this result may be inaccurate to validate other segmentation techniques. A better solution therefore lies in defining a reference segmentation which is algorithm-independent (ground truth). This reference segmentation is manually defined and cross-checked by several people using the same set of rules to segment the video. In order to obtain consistent results, the definition of a transition effect must be given in the clearest and most unambiguous possible way.

3.1 Definitions of Transition Effects

Common definitions used in the technical field of audiovisual production should be used to validate the segmentation algorithms. Currently, three types of transition effects have been used which are called "cuts", "dissolves" and "others". These types may also be sub-divided in order to obtain a finer classification. For instance, the class of "dissolves" effect may be split into "dissolve", "fade in" and "fade out" effects. The effects may also be labelled according to their semantic importance to account for the context in which they appear.

3.2 Rules for Validating the Transition Effects

Precise rules have to be defined for the various effects and their possible subdivisions. "cuts", "dissolves" and "others" effects have an intuitive significance though their exact technical definition should be much more precise and sometimes complex, and may also depend on the target application. Experience shows that giving such formal definitions is not always straightforward. For instance, it has been decided in our experimental works that temporal discontinuities within "visual jingles" and stroboscopic effects should not be counted as valid cuts. In our context, a transition is counted only if it corresponds to a transition which applies to the complete frame. Using this definition, superimposed text, small images and logos appearance and disappearance are therefore not counted as transition effects.

4 Comparison with the Reference Segmentation

In order to evaluate the reliability of a technique, one needs to compare the segmentation given by the system with the reference segmentation according to given rules. These rules must define the error and correctness in the specific context of the considered application. For example, a reference segmentation may describe cuts, fades, dissolves and wipes and the segmentation algorithm only be able to detect cuts and fades without distinction between fades and cross-dissolves.

For each individual type of effects (“cut”, “dissolve” or “other”) and for any of their combinations, we need to count the insertions N_{I} (false detections) and the deletions N_{D} (missed effects) between the reference indexation and the indexation produced by the tested system. For this, additional rules (distinct from the ones used to determine whether an effect is actually present or not) must be defined in order to determine whether an effect is correctly matched between a segmentation and the other. For instance, it may be decided that a dissolve effect has been correctly detected if and only if it partly overlaps by at least 50 % with the correct one, in order to allow for approximate boundary detection (and/or indexation). Also it must be decided if a confusion between a “dissolve” and an “other” or between a “cut” and a very short “dissolve” should be counted as an error or not. Such rules, parameterisable in order to take into account the application field specificities, have been implemented in a computer program that is able to automatically and deterministically count the number of insertions and deletions for each type of effect.

The total number N_{T} of effects of each type as well as the total number N_{F} of frames in the document (or database) must also be counted in order to be able to compute the various quality criteria.

5 Selection of “Quality Measure” Criteria

Different measures have been proposed to compute the error or success rate over the results of different segmentation methods. For a given formula that computes the error rate (eg. the number of inserted transition effects over the number of real ones) the results vary significantly and depend on the definitions of error and success within the application domain. We recall below some examples of the formulas that have already been proposed to evaluate temporal segmentation methods.

5.1 Accuracy

A simple expression to compute the accuracy is proposed by Aigrain and Joly in [7] which is also equivalent to a measure commonly used for the evaluation of speech recognition systems [4] :

$$\mathrm{Accuracy} = \frac{N_{\mathrm{T}} - (N_{\mathrm{D}} + N_{\mathrm{I}})}{N_{\mathrm{T}}} = \frac{N_{\mathrm{C}} - N_{\mathrm{I}}}{N_{\mathrm{T}}}, \qquad (1)$$

where N_{T}, N_{D}, N_{I} and N_{C} are respectively the number of actual transition effects present in the video database, the number of transition effects deleted, inserted and correctly found by the tested system.

Counter-intuitive results may be obtained using this measure (Accuracy < 0) when $N_{\mathrm{I}} > N_{\mathrm{C}}$ or $(N_{\mathrm{D}} + N_{\mathrm{I}}) > N_{\mathrm{T}}$. This may happen when the number of errors is larger or equal than the number of transitions. Moreover, it is important to include the size of the video sequence in the evaluation of a segmentation method since the number of errors may potentially be equal to the number of frames N_{F}.

5.2 Error Rate

The previous measure do not take into account the complexity of the video sequence nor its size. Corridoni and Del Bimbo[3] propose a measure that evaluates the error rate (insertion and deletion of transition effects) over the whole results of the segmentation algorithm:

$$\text{Error Rate} = \frac{N_{\mathrm{D}} + N_{\mathrm{I}}}{N_{\mathrm{T}} + N_{\mathrm{I}}} = \frac{N_{\mathrm{D}} + N_{\mathrm{I}}}{N_{\mathrm{C}} + N_{\mathrm{D}} + N_{\mathrm{I}}} \tag{2}$$

Here again, this measure does not include the complexity and size of the test video sequence. Moreover, this measure is not adequate for the evaluation and the comparison of methods because it implicitly gives more importance to deleted transition effects than to inserted ones. This importance is not weighted with an explicit factor and is therefore difficult to assess. For example, for a video sequence containing 10 transition effects, we obtain Error Rate $= \frac{1}{3}$ if the segmentation technique produces 5 effect insertions (i.e. $N_{\mathrm{T}} = 10, N_{\mathrm{D}} = 0, N_{\mathrm{I}} = 5$). By contrast, the error rate increases (Error Rate $= \frac{1}{2}$) in the case of 5 deletions (i.e. $N_{\mathrm{T}} = 10, N_{\mathrm{D}} = 5, N_{\mathrm{I}} = 0$).

5.3 Recall and Precision

The measure used by Boreczky and Rowe [2] can be applied in different contexts. They use the Recall (which is the ratio between desired found items), and the Precision (which is the ratio of found items that are desired).

$$\text{Recall} = \frac{N_{\mathrm{C}}}{N_{\mathrm{C}} + N_{\mathrm{D}}} \qquad \text{Precision} = \frac{N_{\mathrm{C}}}{N_{\mathrm{C}} + N_{\mathrm{I}}} \tag{3}$$

The results produced by these formula are not normalised and therefore difficult to compare one to another. In this respect, graphs of the Recall are displayed as a function of the Precision for different threshold values. Different Recall values are given for a given Precision value since these measures are, in general, compensated: if the evaluated segmentation method is very strict, the number of deleted transition effects increases while the number of inserted effects decreases. The consequence is a decreasing Recall value against an increasing Precision value.

These two above parameters are strongly correlated so that their global evaluation shows the same problems as in the previous measures.

5.4 Time Boundary and Classification Errors

Hampapur and Jain have proposed a very interesting application-oriented measure [6]. They consider the following two types of errors in the detection of transition effects: the type of the transition effects recognised and the temporal precision of the segmentation. One can increase the weight corresponding to a given error type according to the segmentation application. To compute the error, their measure compares the results of the automated segmentation to those obtained with a manual segmentation (which is supposed to be the reference, containing only correct information). The results are therefore made more reliable at the cost of an extra hard (and tedious) work during the phase of manual segmentation.

$$E(V, V') = E_{\mathrm{LS}} * W_{\mathrm{LS}} + E_{\mathrm{SC}} * W_{\mathrm{SC}}, \tag{4}$$

where $V : \{S_1, S_2, ..., S_N\}$ is the manual video segmentation in N segments (S_n), $V' = \{S'_1, S'_2, ..., S'_K\}$ is the automated video segmentation, E_{LS} is the error in terms of segment temporal limit defined by the transition effects, W_{LS} is the weight of the temporal limit error regarding to the application, E_{SC} is the error of mis-classification of transition effects and W_{SC} is the weight of the classification error.

To compute the error, the segments are matched one to another and the maximal overlap between corresponding segments in the two videos is computed.

5.5 Definition of a Common Description Format

Different file formats have been used by the authors to store the results of the segmentation process. Several hard transformations have been required to build a unique reference segmentation from independent manual segmentations. We used the straightforward text (`txt`) format in a first approach but there exist formats which are more suited to the description. Such formats include `rdf` or `xml` or some other proposals made by the MPEG-7 group [9].

5.6 Performance Measure

The performance measure proposed by Hampapur and Jain [6] is related to the application domain and is able to compute the errors made by most of segmentation methods. However, the boundary segment error is very difficult to evaluate in the case of "dissolve" effects. The manual segmentation is therefore not reliable in this case. We propose here some measures which overcome these shortcomings.

Error Probability This measure computes the probability to make an error (deletion or insertion) when an error is possible. The temporal segmentation methods can make a detection error on each video frame.

$$P(e|ep) = \frac{N_{\mathrm{D}} + N_{\mathrm{I}}}{N_{\mathrm{F}}} \tag{5}$$

Insertion Probability The insertion probability is the probability that a transition effect is detected where no effect is present.

$$P(\text{insertion}) = P(\text{detection}|\text{no effect}) = \frac{N_{\mathrm{I}}}{N_{\mathrm{F}} - N_{\mathrm{T}}} \tag{6}$$

Deletion Probability It is the probabilty of failing to detect an effect when the effect exists.

$$P(\text{deletion}) = P(\text{no detection}|\text{effect}) = \frac{N_{\text{D}}}{N_{\text{T}}} \tag{7}$$

Correctness Probability It is the probability to detect a transition effect when it exists and not to detect it when it does not exists. One can give more importance to either of these situations by using a weigth $(k1, k2)$.

$$P(\text{correction}) = k1 * P(\text{detection}|\text{effect}) + k2 * P(\text{no detection}|\text{no effect}) =$$

$$k1 * (1 - P(\text{deletion})) + k2 * (1 - P(\text{insertion})) =$$

$$(k1 + k2)(1 - (k1' * P(\text{deletion}) + k2' * P(\text{insertion}))) \tag{8}$$

Where $k1, k2, k1'$ and $k2'$ take values between 0.0 and 1.0 but (k1+k2=1). These measures take into account both the total number of transition effects (N_{T}) and the total number of frames of the sequence in question (N_{F}). This makes these measures more robust to the problems encountered using the previous definitions. For the results showed in the last section we have used $k1 = k2 = k1' = k2' = 0.5$.

5.7 Method Complexity Evaluation

In order to compare the temporal segmentation methods it is important to use a measure of complexity in relation to the computing time induced, the need for learning and the threshold measuring the dependence on the document under investigation.

5.8 Complexity of the detection of a transition effect

The detection of a transition may not be consistently difficult. The complexity of this detection must therefore be included within the evaluation process in order to weight the possible type of errors made. Examples of such evaluations are as follows.

- Maximum and Minimum Difference between Histograms. If the difference between the frames located before and after an effect is above a certain threshold, the detection of the transition is easy and an error made at this location can be considered as serious. On the other hand, if this difference is large between two frames within a shot (e.g. flash), a transition effect insertion is likely to be made at this place.
- Motion Quantities:
  - Histogram of the Spatial Derivative Difference: a difference exceeding a given threshold may denote the precence of motion or camera work, and a transition effect insertion at this location is likely to happen.
  - Histogram of the Difference between Boundary Phase.

  - Autocorrelation measure.

- Insertion and Deletion Probabilities. This therefore results in defining the probability of the insertion of a transition effect over the histogram difference and the probability of correct detection over the histogram difference, respectively given as,

$$P(\text{insertion}|\mathit{\Delta}\text{H}) \text{ and } P(\text{correct detection}|\mathit{\Delta}\text{H}).$$

6 Temporal Video Segmentation Methods

Many automated tools for the temporal segmentation of video streams have been already proposed. It is possible to find some papers that are providing state of the art of such methods (see e.g. [5]). We briefly present in this section the various systems that have been used in our comparative evaluation.

6.1 The "LIMSI" system

The "LIMSI" system was developed at the LIMSI-CNRS laboratory and was then improved at the CLIPS-IMAG laboratory [8]. It detects "cuts" by direct image comparison after motion compensation and "dissolves" by comparing the norms of the first and second temporal derivatives of the images. This system also includes a special feature for detecting photographic flashes and filtering them as erroneous "cuts".

6.2 The "CLIPS" system

The "CLIPS" system was developed at the CLIPS-IMAG laboratory [8]. It only detects "cuts". It uses two separate detection subsystems respectively based on colour histogram comparison and on rough edge tracking, and then merges the results.

6.3 The "IRIT-LIP6" system

The algorithm was developped at the IRIT [7] but has been improved to speed up the computation step at the LIP6 (3/4 times of the real time, with a soft mpeg decoding on a pc 233Mhz). It detects both "cuts" and "fades" and includes some filtering functionalities to deal with photographic flash effects, fast motion and dark scenes.

7 Results

The reference files are based on the manual validation made by the authors with the contribution of the INA.

Table 1 gives the error rates with respect to the methods. Each given value corresponds to a mean of the error rates obtained on each document of the corpus. Since not all the tested methods are able to handle every possible type of effect, only cuts were used in our evaluation. The methods noted "Method #" are classic segmentation methods. From "Meth 1" to "Meth 9" respectively, the methods in table 1 are: histogram difference, intensities difference, difference between the addition of intensities, histogram intersection, invariants moments difference, thersholded intensities variation, correlation rate, χ^2 formula, χ^2 formula over the blocks.

All values are presented in percentage.

Table 1. Several reliability measures of cut segmentation.

Method	Accur.	Prec.	Recall	Error R.	Ins. Pr	Del. Pr	Error Pr	Correc. Pr
LIMSI	76.1	83.18	95.4	20.00	0.17	4.57	0.21	97.62
CLIPS	81.7	88.76	93.6	16.35	0.10	6.44	0.16	96.72
LIP6	63.0	82.67	79.6	31.76	0.15	20.38	0.33	89.73
Meth 1	-157.2	24.88	78.0	76.75	2.11	22.03	2.29	87.92
Meth 2	37.0	65.48	78.1	44.67	0.37	21.90	0.56	88.86
Meth 3	77.8	90.85	86.5	20.39	0.07	13.45	0.19	93.23
Meth 4	-27.9	41.67	69.9	64.66	0.87	30.09	1.14	84.51
Meth 5	-38.5	40.88	86.5	61.55	1.12	13.45	1.23	92.70
Meth 6	-373.3	16.44	91.5	83.80	4.18	8.49	4.22	93.66
Meth 7	1.2	50.29	95.3	50.90	0.84	4.66	0.88	97.24
Meth 8	-137.2	21.92	53.6	81.56	1.71	46.38	2.11	75.94
Meth 9	-1462.5	3.50	55.2	96.59	13.65	44.81	13.93	70.76

Table 2 details the results obtained using the techniques developped by the authors of this paper.

Table 2. Mean reliabilty measures.

Method	Type	Accur.	Prec.	Recall	Error R.	Ins. Pr	Del. Pr	Error Pr	Correc. Pr
LIMSI	Cuts	76.1	83.18	95.4	20.00	0.17	4.57	0.21	97.62
	Diss.	21.9	64.96	46.9	62.76	0.01	53.40	0.05	73.28
	Total	70.0	82.28	89.2	25.78	0.19	10.82	0.29	94.48
CLIPS	Cuts	81.7	88.76	93.6	16.35	0.10	6.44	0.16	96.72
	Diss.	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Total	73.8	88.81	84.4	24.28	0.10	15.55	0.26	92.16
LIP6	Cuts	63.0	82.67	79.6	31.76	0.15	20.38	0.33	89.73
	Diss.	-63.5	21.91	25.5	86.74	0.06	74.86	0.12	62.53
	Total	51.9	77.25	73.6	40.44	0.21	26.42	0.47	86.67

Finally, table 3 shows gives the method performances for each document in the corpus.

Table 3. Correctness probability when detecting cuts, dissolves and both for each document and method.

Method	Type	aim1	aim2	aim3	aim4	aim5	aim6	aim7	aim8
LIMSI	Cuts	99.78	96.61	97.52	96.47	91.70	99.60	99.33	98.88
	Diss.	72.44	76.34	83.31	65.97	66.66	67.84	74.99	95.44
	Total	95.04	92.34	97.22	89.91	88.27	96.44	94.21	98.68
CLIPS	Cuts	99.61	93.70	97.04	95.97	91.28	99.25	98.07	98.80
	Diss.	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Total	91.23	87.63	96.11	88.44	87.31	94.54	92.03	97.89
LIP6	Cuts	98.84	65.29	95.50	95.22	86.97	96.81	97.46	98.36
	Diss.	65.27	57.16	49.97	57.95	60.40	82.11	69.96	63.61
	Total	92.96	63.62	94.58	88.09	83.50	95.49	93.66	97.65

8 Conclusions

This work has shown that objective comparison between different temporal video segmentation systems is feasible using a common corpus, a corresponding reference segmentation of it, an automatic comparison tool of document segmentations, and an appropriate global "quality criterion".

This work must be completed by a more accurate definition of transition effects. It should also be extended to some other types of contents to be able to deal with other classes of temporal segmentations (speaker segmentation, camera work segmentation, and so on).

The main difficulty resides in reaching a consensus on a common definition of transition effects, a common reference file format and a common mean of evaluation. The experience acquired in the speech recognition domain should be used as a guidance and periodic comparative performance tests should similarly be set up for evaluating temporal video segmentation systems.

References

1. G. Ahanger and T. D. C. Little.: A survey of technologies for parsing and indexing digital video. Journal of Visual Communication and Image Representation. **7** (1996) 28–43
2. J. S. Boreczky and L. A. Rowe.: Comparison of video shot boundary detection techniques. Proceedings Spie - The International Society for Optical Engineering. (1996) 170–179

3. J. M. Corridoni and A. Del Bimbo.: Film semantic analysis. CAIP95. Czech Republic. (1995)
4. Proceedings of DARPA Speech and Natural Language Workshop, January (1993)
5. F. Idris and S. Panchanathan.: Review of image and video indexing techniques. Journal of Visual Communication and Image Representation. **8** (1997) 146–166
6. Hampapur, A., Jain, R., Weymouth, E.: Production Model Based Digital Video Segmentation. Multimedia Tools and Applications **1** (1995) 9–46
7. Aigrain, P., Joly, P.: The automatic real-time analysis of film editing and transition effects and its applications. *Computers and Graphics* **18**(1) (1994) 93–103
8. G. M. Quénot and P. Mulhem. Two Systems for Temporal Video Segmentation, Submitted to Content Based Multimedia Indexing, Toulouse, Oct. (1999)
9. ISO-IEC JTC1/SC29/WG11/N2467 Description of MPEG-7 Content Set http://drogo.cselt.stet.it/mpeg/public/w2467.html