Poster proposal for ACM MM 98

Using content models to build audio-video summaries

Janne Saarela, Bernard Merialdo
Institut Eurécom
BP 193, 06904 Sophia-Antipolis, FRANCE
{saarela,merialdo}@eurecom.fr

Introduction

In this presentation, we describe ongoing work towards the definition of a theoretical model for generating multimedia summaries. Our motivation is based on two assumptions:

- there is no ideal summary. Based on what the user is looking for, different summaries could be generated in order to best fit user's needs.
- the use of a theoretical model should allow to formalize (at least some of) the various parameters that can be introduced in a summary, and to facilitate the construction of an optimal summary in each case.

Our current work aims at understanding the various types of contents that can be defined in an audio-video sequence, clarify the relationship between audio and video segments, and experiment with different types of summarisation processes. This presentation focuses on two aspects: the study of synchronisation constraints between audio and video, in particular synchronisation, and a proposal for a first model for measuring video content.

Constructing video summaries

We consider the usual paradigm of summarisation (also known as skimming) which constructs the summary by concatenating pieces of the original video. As described in [Mayb97], this process is a sequence of steps:

analysis -> selection/condensation -> presentation

Automatic analysis is technically the most difficult step to implement, because it involves using advanced pattern recognition techniques on problems that are not yet completely solved. We focus on a model for automating the selection process.

We feel that there is no ideal summary. Based on the type of document (movie, documentary, news program...), on user preferences (style) and more importantly on user's intentions (knowing the end of a movie, getting all places in a documentary...), the best summary may be very different. A summary will be used in a "preview" mechanism so that the user can quickly decide if he wants to get/buy/watch the whole video. Depending on the kind of information that the user is looking for, the summary should contain different elements, and the selection criteria should be different.

We propose to use models where the audio and video tracks are segmented into units (e.g. using pauses in the audio and cut detection in the video). Each segment is assigned an importance value (e.g. using TD-IDF model for the audio). The summarisation process is then a problem of constrained optimization, trying to select segments with highest value, while respecting certain constrains in the selection.

Constraints

Many constraints can be considered. Following are what we feel are the most important among the ones we have considered:

- minimum shot duration: short segments are difficult to grasp and required tremendous mental effort from the user. Minimum duration could be taken as 3 to 5 seconds [Chri98].
- audio-video synchronisation: in certain cases, audio and video should be synchronised (the

question of whether this should always be the case is investigated in the next section).

- reordering: should segments be presented in the same order in which they appear in the original sequence.
- redundancy: multiple occurences of a segment (or similar) can be detected, and only one introduced in the summary. Also, sometimes an excerpt of a segment can be as informative as the segment itself.

These constraints define the set of admissible summaries, among which the optimization procedure should choose the most informative one. They have to be carefully studied because if a too strong constraint is enforced, this will reduce the value of the maximum found by optimization.

Audio-Video synchronisation

In [Chri98], it is found that users find audio-video synchronisation a desirable feature in summaries. This is a strong constraint and causes a waste of efficiency since interesting video segments are sometimes associated with uninteresting audio segments. In order to further investigate whether non-synchronised audio and video segments could be used in summaries, we designed an experiment where we present the user with randomly chosen audio and video segments and let the user guess whether they are synchronised or not. The errors rates over 4 users are reported in the following table:

% errors in guessing synchronization (mean,variance)	
all segments	25%, 4.12
person segments	17%, 3.32
non-person segments	33%, 5.20
synchronised segments	23.75%, 12.44
non-synchronised segments	25.83%, 7.59

While the overall error rate in guessing synchronisation is 25%, we can notice that segments containing persons provide a lower error rate (bad synchronisation is easier to detect), and non-person segments a higher error rate (bad synchronisation is more difficult to detect). When we compare synchronised versus non-synchronised segments, we find that the error rate is about the same, with a large variance. These results suggest that presenting synchronised segments is not an absolute requirement, but rather a constraint based on the content of the segments, so that certain segments could be included into a summary with a non-synchronised audio part without disturbing users.

A model for video content

We propose a model for video content which aims at providing a quantitative measure of the content of a segment with respect to the whole video, and is inspired by the TF-IDF model for natural language. Many variations are possible, but we only present a simplistic case to illustrate our approach.

Each video segment is evaluated with respect to three attributes: people (important persons which appear in the shot), action (what is happening) and location (where the shot was taken). We make the drastically simplifying assumption that each segment can be associated with one unique value for each attribute, so that a segment can be seen as a triplet (people, action, location) and a duration.

As an example, we have hand-labelled a 50 min. video (documentary, 240 segments), to provide the following results (attributes, number of values and most frequent values):

People (21 values)	Action (16 values)	Location (20 values)
<none> 40.86	Discussion 31.71	Moscow 16.89
Engineer 12.67	Narration 30.06	Home 15.60
Joe 8.67	Landscape 9.59	Laboratory 9.17
Robert Kennedy Jr. 4.79 ...	<none> 5.04 ...	New York 8.46 ...

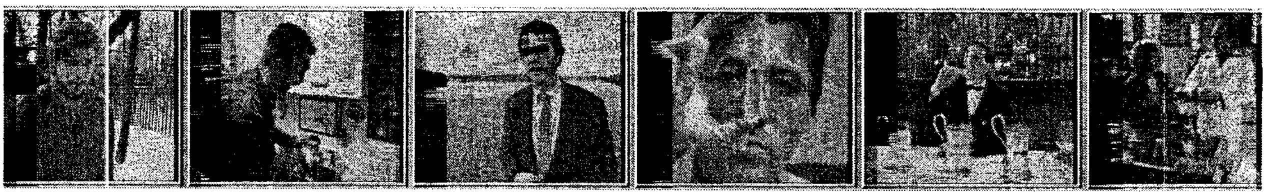
(Note that we can hope to be able to automate some of this labelling in the future by using some image processing: locations as backgrounds, people as foreground objects, actions as movements). The probability of an attribute value is the percentage of time this value is present in the video. We can define a content model for a segment $s = (p_s, a_s, l_s)$ using the following formula to compute the content value:

$$P(s) = \alpha P(p_s) + \beta P(a_s) + \gamma P(l_s)$$

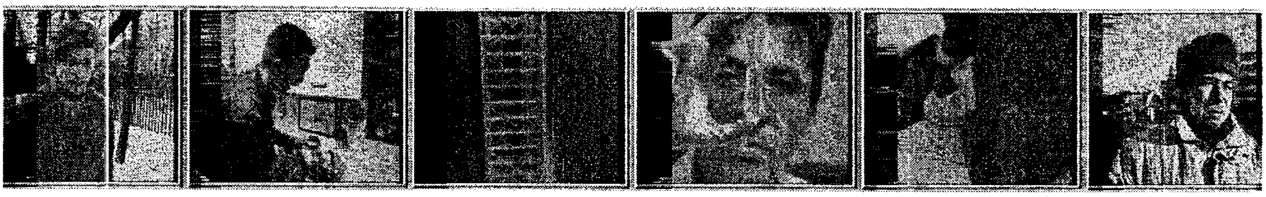
where the coefficients α, β, γ are positive and sum to 1. Varying these coefficients allow to give more or less importance to each attribute.

The following pictures show the best shots selected according to video only (no audio) using various coefficients:

- emphasis on people



- emphasis on location



(note that some shots are selected by both methods).

References

[Chri98] Christel, M.G., Smith, M.A., Taylor, C.R, & Winkler, D.B. Evolving Video Skims into Useful Multimedia AbstractionsVideo abstracting. *Proceedings of the CHI '98 Conference on Human Factors in Computing Systems*, April, 1998

[Lien97] Rainer Lienhart, Silvia Pfeiffer, and Wolfgang Effelsberg. Video abstracting. *Communications of the ACM*, 40(12):54–62, December 1997.

[Mayb97] Mark Maybury and A Merlino. Multimedia summaries of broadcast news. *International Conference on Intelligent Information Systems*, December 1997.

[Smit97] Michael Smith and Takeo Kanade. Video skimming for quick browsing based on audio and image characterization, IEEE *CVPR*, February 1997.