Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 111 (2008) 142–154
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Audiovisual integration with Segment Models for tennis video parsing

Manolis Delakis [a,*], Guillaume Gravier [b], Patrick Gros [c]

[a] IRISA/University of Rennes 1, Campus de Beaulieu, 35042 Rennes, France
[b] IRISA/CNRS, Campus de Beaulieu, 35042 Rennes, France
[c] IRISA/INRIA, Campus de Beaulieu, 35042 Rennes, France

Received 27 December 2006; accepted 10 September 2007
Available online 22 September 2007

Abstract

Automatic video content analysis is an emerging research subject with numerous applications to large video databases and personal video recording systems. The aim of this study is to fuse multimodal information in order to automatically parse the underlying structure of tennis broadcasts. The frame-based observation distributions of Hidden Markov Models are too strict in modeling heterogeneous audiovisual data. We propose instead the use of segmental features, of the framework of Segment Models, to overcome this limitation and extend the synchronization points to the segment boundaries. Considering each segment as a video scene, auditory and visual features collected inside the scene boundaries can thus be sampled and modeled with their native sampling rates and models. Experimental results on a corpus of 15-h tennis video demonstrated a performance superiority of Segment Models with synchronous audiovisual fusion over Hidden Markov Models. Results though with asynchronous fusion are less optimistic.

Keywords: Hidden Markov Models; Segment Models; Multimodal fusion; Video indexing; Video summarization

1. Introduction

Automatic annotation of video documents is a powerful tool for managing large video databases and, more recently, for the development of sophisticated consumer products that meet high-level user needs like highlight extraction. One can accomplish this task by using explicit hand-crafted and thus domain-dependent models which can perform reasonably well in some cases [1]. However, more effective ways are needed to bridge the required high-level user needs and the low-level video features at hand, such as image histograms or speaker excitation. A key question towards this end is an efficient video content representation scheme [2]. Hidden Markov Model [3] (HMM) is a powerful statistical approach for modeling video content and can be used as a statistical parser of a video sequence [4], sharing notions from the field of speech recognition.

We use Markovian models for tennis broadcasts structure analysis. In this type of video, game rules as well as production rules result in a structured document. In a previous work, we used flat or hierarchical HMMs [5] to parse this structure and to segment raw video data into human-meaningful scenes. The table of contents of the video can then be automatically constructed.

Audiovisual integration with HMMs is generally addressed with a concatenative fusion scheme that assumes homogeneous and synchronous features. However, the visual and auditory modalities are sampled at different rates. In addition, the visual content follows the production rules while the auditory one captures raw sounds from the court, interlaced with commentary speech. There is thus, firstly, a certain degree of asynchrony between auditory and visual features and, secondly, they follow different temporal models.

In this study, we introduce video indexing with Segment Models [6] (SMs) as a means of a more efficient and

* Corresponding author.
E-mail addresses: Manolis.Delakis@irisa.fr (M. Delakis), Guillaume.Gravier@irisa.fr (G. Gravier), Patrick.Gros@irisa.fr (P. Gros).

1077-3142/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2007.09.002

versatile multimodal fusion and provide an experimental comparison with HMM-based fusion. With SMs, the synchrony constraints between the modalities can be relaxed to the scene boundaries, thus enabling to process each modality with their native sampling rates and models. The aim of this work is the exploration of the many possibilities of audiovisual integration that SMs can offer rather than the design of an integrated and general-purpose tennis video parser. To this end, we employ a small but sufficient to our needs set of audiovisual features and restrict ourselves to the producer styles of French television.

This paper is organized as follows. Relative work on the HMM-based multimodal fusion is given in Section 2. Ground definitions on the problem at hand are provided in Section 3. Feature extraction is discussed in Section 4. In Section 5 we see how the visual content is modeled by HMMs and SMs, stressing conceptual differences between these two modeling alternatives. Audiovisual integration is then discussed in Section 6. Parameter estimation details are provided in Section 7 and experimental results in Section 8. Finally, Section 9 concludes this study.

2. Previous work on multimodal fusion

For a detailed review of the various aspects of the multimodal video indexing problem, the interested reader is referred to the literature reviews [1,7,2,8]. In this section we will focus on HMMs, which are widely used to exploit the temporal aspect of video data. Indeed, depending on the video genre and the production rules, video events occur with a temporal order that will finally reveal the semantics. HMMs provide a powerful statistical framework for handling sequential data and they are thus a natural candidate for learning temporal dependencies in video. Wolf [4] introduced HMMs as a *statistical parser* of the syntax of a video document. Low-level video features are considered as the observations of a hidden Markovian stochastic process that represents the video syntax. The Viterbi algorithm is then employed to recover the syntax, given the video model (*i.e.* the observation and transition probabilities).

A simple scenario of multimodal integration with HMMs is to perform classification or indexing based on each modality independently and then to fuse likelihoods in order to take the final decision. This fusion scheme, referred to as *late fusion*, was proposed in [9] among others.

The second approach to HMM-based multimodal integration is *early fusion*, *i.e.* the concatenative fusion of features collected from all the modalities into a super-vector of observations. This fusion scheme is a popular choice in the video indexing community due to its simplicity, and has been used for video segmentation [10,11], human dialog detection [12], and TV broadcasts classification [9,7,13–15] on top of visual and auditory features. Early fusion has also been widely studied in the field of audiovisual speech recognition [16]. However, concatenative fusion requires to convert the sampling rates of each modality, which is always data dependent. The concatenation of visual features (usually sampled at 25 fps) to the auditory ones (usually sampled at 100 fps) can be done by interpolation [16], and the opposite conversion by averaging [11]. Nevertheless, in many video indexing approaches, features from large video segments (like a video shot) are collected and then pre-classified to yield descriptors. In so doing, the problem of the sampling rate conversion is artificially bypassed.

The underlying assumption however of the early fusion scheme is that all the modalities should be synchronous and should follow the same model topology, which does not generally hold. In order to remedy this forced state synchrony between the modalities and to increase modeling capabilities, a number of HMM variants have been proposed in the audiovisual speech recognition and video indexing communities. In multistream HMMs [17], each modality (or stream) is modeled by independent HMMs which are forced to synchronize at some fixed points. In synchronous multistream HMMs, the states themselves are these points. In practice, these models do not differ with the state synchronous HMMs, except for the explicit assumption of conditional independence of the observation streams. This scheme allows for the incorporation of state-dependent weighting or reliability measurements on each modality [18].

In asynchronous multistream HMMs, the synchronization points are extended beyond the hidden states, like the end of phonemes in audiovisual speech recognition. Between these points, the streams are considered independent and are modeled by individual unimodal HMMs. The key idea is that the unimodal HMMs can follow different topologies and also operate at the native sampling rate of their modality. The likelihoods (or scores) produced by the unimodal HMMs are recombined at the synchronization points with probability products or with any other recombination function. For practical reasons however, during decoding asynchronous multistream HMMs have been used in the form of a product HMM in audiovisual speech recognition [16]. In practice, product HMMs still require same topology and sampling rates across the modalities, while they can handle some limited asynchrony because, in order to keep model complexity low, not all the possible product-state transitions are allowed. In addition, when more than two modalities are fused, the overall model complexity becomes intractable.

The Asynchronous Hidden Markov Model (AHMM) [19] is a special HMM architecture designed to jointly model pairs of lightly asynchronous streams and containing different number of samples. AHMMs process the two streams by letting the shorter stream be stretched in time in order to meet a better match with the longer one. During Viterbi decoding, the search is performed not only through all the possible hidden state paths but also through all the possible alignments between the two streams. AHMMs were introduced in audiovisual speech recognition, in order to account for noisy desynchronization of