Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 29 (2008) 181–191
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Parametric model for video content analysis

Wei Chen *, Yu-Jin Zhang

Department of Electronic Engineering, Image Engineering Laboratory, Tsinghua University, Beijing 100084, China

Received 8 April 2006; received in revised form 12 July 2007
Available online 24 October 2007

Communicated by H.H.S. Ip

Abstract

In this paper, we propose a parametric model for the video content analysis. We use the autoregressive (AR) modeling to model the frame feature sequence over time and make the future analysis in the AR parametric space. Based on our parametric framework, applications of detecting shot boundaries in video sequences, extracting key frames and combined spatial–temporal features for shot classification are proposed. Experiments show that our new parametric framework can present the video content better than the traditional color histogram.

Keywords: Parametric model; Shot boundary; Key frame; Spatial–temporal feature

1. Introduction

Video content analysis has been an active research area in recent years. The video signals which can be viewed as an image sequence have both temporal and spatial features. Many combined spatial–temporal features have been applied to this field (Hicham et al., 2005; Laptev and Lindeberg, 2004; Ngo et al., 2002). These spatial–temporal approaches are powerful for many applications of video content analysis. In (Coudert et al., 1999), a 1D representation of video frames by a Mojette transform is proposed. This feature is successfully used for the applications of motion estimation, scene change detection and areas of interest extraction. In (Nicolas et al., 2004), shots are grouped by employing 1D mosaics based on X-ray projections of the video frames, which represent the integration along vertical and horizontal axes. The crucial problem for these works is how to choose proper spatial and temporal descriptors according to their special applications.

For the video applications such as shot classification, video indexing and video retrieval, the temporal information is more significant. To model the temporal relations of the spatial feature sequence, several time-series modeling algorithms have been already taken to this field. One common method of the temporal characterizing is the probability models with Bayes learning. For such models, both the knowledge from the concepts and mining in the data can be merged and learned in a Bayes framework. In (Xu et al., 2005), a hidden Markov model (HMM) multi-level framework is proposed. Semantics in different granularities are mapped to a hierarchical model space. If the models and prior distributions lead to intractable posterior distributions, numerical integration methods can be used to make the answers more accurate, such as the Markov chain Monte Carlo (MCMC). In (Zhai and Shah, 2006), the shots are modeled into scenes by calculating the posterior probability of the scenes number and their corresponding locations in the way of sampling from their target distributions with the MCMC technique. However, there are also several drawbacks of such approaches. With the specified distribution forms and the prior parameter distributions relying on the knowledge of a specific domain (for example, a particular sport game), the application scope of the

* Corresponding author. Tel.: +86 10 62781291; fax: +86 10 62770317.
E-mail addresses: chenweith00@mails.tsinghua.edu.cn (W. Chen), Zhang-yj@mail.tsinghua.edu.cn (Y.-J. Zhang).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.09.020

HMM model is restricted. The numerical integration methods would also cost great computational resources to get the accurate results.

Another way of characterizing temporal relations of the video signal is employing direct temporal features, such as the optical flow (Laptev and Lindeberg, 2004) and the motion vectors (Rajesh and Michael, 2002; Duan et al., 2005). In (Laptev and Lindeberg, 2004), the authors use a local descriptor to present interested regions and its corresponding temporal information by the optical flow. In (Rajesh and Michael, 2002), the motion fields are treated as a distinct signal analogous to time-series and a mechanism to filter the motion fields is present. In (Ngo et al., 2002), the spatial–temporal slices are used to present motion patterns and the key-frame is extracted mainly by the motion patterns. However, most of those features lack the ability to measure long range temporal relations, and the relations in such features are hard to quantify.

In this paper, we employ a framework to model the temporal information of the video sequence in a universal parametric space. If such a model could be learned either from the data or from the physics of the actual scenario, it would help significantly in problems such as identifying and synthesizing video sequences. We employ a time-series parametric model here, namely the ARMA model, not relying on the specific distribution but reflecting the underlying temporal relation of the frame sequences. It is well-known that autoregressive (AR), moving-average (MA) and auto-regressive-moving-average (ARMA) models are useful time-domain models for the representation of discrete-time signals (Martin, 2000; Rajesh et al., 1997), especially as parametric methods to estimate the covariance and the power spectral density of the stochastic processes.

With the great ability to present the temporal relation in the frame's spatial feature sequence, this framework is applied to several applications of video content analysis. Our work is mainly on the motivation of shot boundary detection, while it can also be applied to key-frame extraction and shot classification.

The remainder of paper is structured as follows. Section 2 gives an interpretation of AR process model with the recursive learning method proper for our applications and the model distance used in the parametric space. We apply this framework to model the spatial features in the video frame and discuss its fundamental relations with motions. In Section 3 we present our parametric frame work for some most widely used applications, while in Section 4 we demonstrate its performance by experiments. We conclude this paper with a discussion in Section 5.

2. Parametric framework for video analysis

We first present a definition of the parametric modeling, which could be generalized to video feature fields, and then show how to estimate parameters of the model and the distance measurement in the parametric space.

2.1. The parametric model

ARMA model is a simple time-series model that has been used very successfully for prediction and modeling. Compared with the previous temporal modeling approaches, there are several advantages of the parametric AR model. First of all, with the ability to present the complex linear system added with white noise, AR model is suitable for the video temporal structure expression. Moreover, some tough problems in the color-time space such as the gradual shot boundary detections have better properties and are much easier to tackle in this new parametric space, just like what we do in the spectral space with a Fourier transform. Besides, as a crucial property for the applications of video retrieval or video classification, the distance of the temporal relation is easy to measure with the AR parameters in a typical definition (Martin, 2000; Cock and Moor, 2000), but may cause great difficulty for the motion field or the optical flow features. The advantage of the ARMA model to the AR model is that it can characterize systems with both poles and zeros while the AR model can be used with poles only. However, an ARMA model may be approximated as a high-order AR model, much simpler than the original one. Therefore, we choose AR model to estimate the video temporal structure.

AR model is a simple time-series model widely used for prediction and modeling. An AR model with order p can be expressed as follows (Martin, 2000):

$$x_n = \sum_{j=1}^{p} a_j x_{n-j} + \eta_n \tag{1}$$

where η_n is the uncorrelated noise of variance σ.

In the z-domain, the system function is (Martin, 2000)

$$\boldsymbol{H}(z) = \frac{\sigma}{\boldsymbol{A}(z)} = \sigma \frac{1}{\sum_{j=0}^{p} a_j z^{-j}} = \sigma \frac{1}{\prod_{i=1}^{p} (1 - \alpha_i / z)} \tag{2}$$

We define $a_0 = 1$ throughout. The AR coefficients a_j constitute the AR part of the model with the α_i known as the poles of the model since it is a rational function of z.

2.2. Learning the parametric model

As we know, the video content changes un-gradually. Therefore, our method must, in principle, capture the dynamics of video sequences locally and be suited for applications in which different local neighborhoods of the video exhibit different dynamics. Consequently, the proper training algorithm for our AR model presented above must have the following properties: recursive convergence speed should be fast enough to ensure that the model parameters can present the current video structure and the parameters should be sensitive to the new signal.

In our proposed framework, we employ the direct-form recursive least squares (RLS) FIR adaptive filter to complete the AR parameters estimating. The RLS procedure