Visual Abstraction of Wildlife Footage using Gaussian Mixture Models

David Gibson Neill Campbell Barry Thomas
Department of Computer Science
University of Bristol
Bristol
BS8 1UB, United Kingdom
{gibson,campbell,barry}@cs.bris.ac.uk

Abstract

In this paper, we present a novel approach for clip-based key frame extraction. Our framework allows both clips with subtle changes as well as clips containing rapid shot changes, fades and dissolves to be well approximated. We show that creating key frame video abstractions can be achieved by transforming each frame of a video sequence into an eigenspace and then clustering this space using Gaussian Mixture Models (GMMs). An iterative process computes a GMM configuration that best clusters the data based on a maximum likelihood threshold. The image nearest to the centres of each of the GMM components are selected as key frames. Unlike previous work this technique relies on global video clip properties and results show that the key frames extracted give a very good representation of the overall clip content. We show that, by using a single threshold, an operator can easily control the number of representative key frames generated. We also demonstrate that clustering in eigen-time space improves the video abstractions in a quantifiable manner and we demonstrate the application of this technique on a database of 307 *clips of wildlife footage containing dissolves, shot changes, fades, pans, zooms and a wide range of animal behaviours.*

1. Introduction

As computational power and storage capacities increase the creation and use of video databases becomes more extensive. The BBC's Natural History Unit in Bristol has a library consisting of hundreds of thousands of hours of wildlife footage on digital beta tape, the world's largest archive of this kind. To use (or reuse) this material currently involves a costly manual process initially requiring an archivist to enter text descriptions and timecodes of the footage into a database. When accessing the footage the desired type of content is searched via a text database which generally gives a number of candidate tapes. These are then sent (often posted around the world) to the program researchers who traverse and view the tapes looking for appropriate content. This process is repeated many times until the desired content, 'look' and 'feel' of the footage is found. The motivation for this work is in the context of investigating techniques to automatically facilitate archiving, retrieval and searching of a digitised version of this wildlife footage.

The initial database created for our work consists of over 100,000 (around 70 minutes) 30-bit YUV uncompressed PAL(601) frames. These constitute 307 clips, generally between 100 and 1000 frames which often contain more than one shot. The clips contain no audio since this is almost always added in post-production for wildlife films. The clips have few post-production editing effects and content ranging from original cellophane film reeds put on tape via a telecine machine to first cut programme and film segments. Given the nature of this database, this paper focuses on the extraction of a small number of key frames from each clip such that these key frames give a useful and meaningful representation of the overall clip content for use by an editor or producer.

1.1. Previous Work

Previous and related work regarding video analysis includes shot detection, content-based video abstraction and video skimming. Many applications require that video is divided up into component shots so as to facilitate further processing such as indexing, key frame extraction and shot content analysis. Techniques including colour histogram, motion and frequency analysis are used to determine when shot changes have occurred, representative key frames are then extracted from individual shots [6, 7, 5]. Video skimming introduces the use of audio and language analysis to generate condensed representations of the original material [8]. In our approach we are not concerned with shot segmentation, just that the visual abstraction generates a small number of informative key frames. This then facilitates fast and useful browsing of a large database of wildlife video footage.

In [9] image histograms and inter-frame motion features have been used as the basis of an eigenspace in order to ob-