Available online at www.sciencedirect.com

SCIENCE DIRECT®

J. Vis. Commun. Image R. 16 (2005) 68–92

JOURNAL OF VISUAL Communication & IMAGE Representation

www.elsevier.com/locate/yjvci

Automatic extraction of moving objects for head–shoulder video sequence

Chung Ming Kuo,* Chaur Heh Hsieh, and Yong Ren Huang

Department of Information Engineering, I-Shou University, Tahsu, 840, Kaohsiung, Taiwan, ROC

Received 30 October 2003; accepted 8 January 2004
Available online 27 February 2004

Abstract

Recently, video object extraction has received great attention because it is a critical technique in object-based video processing. This paper presents a temporal-to-spatial segmentation technique to extract object from a video sequence. The temporal phase employs a simple blockwise temporal-activity measure to approximately locate the object boundary. And then a block-based maximum a posteriori (MAP) scheme, which exploits spatial features of image blocks around the approximated boundary, is adopted to refine the temporal segmentation result. The proposed technique achieves good segmentation quality with very low computational cost for head-and-shoulder sequences with static background.

Keywords: Object-based; Video object; Segmentation; Temporal-activity; MAP

1. Introduction

In the past few years, the idea of object-based video coding has been proposed to achieve high compression efficiency and to allow for more multimedia functionalities. Recently, a new objected-based coding standard, MPEG-4 (Chiariglione, 1997; ISO/IEC, 1998; Koenen et al., 1997; MPEG Video Group, 2001), has been released for multimedia applications. MPEG-4 standard provides users a new level of interaction with visual contents. It offers a framework to view, access and manipulate *objects* rather than pixels, with great error robustness at a large range of bit rates. Obviously, to achieve the object-based video coding, the video sequences are needed

* Corresponding author. Fax: 886-7-6578944.
E-mail address: kuocm@isu.edu.tw (C.M. Kuo).

1047-3203/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2004.01.004

to decompose into individual objects or so called video object plane (VOP), and each VOP is then coded independently. The VOP extraction is a crucial issue for the object-based video coding (Alatan et al., 1998; Salembier and Marqu'es, 1999).

The extraction of video objects may be done by the existing video segmentation techniques, which are categorized into temporal-to-spatial (Diehl, 1991; Hotter, 1990; Musmann et al., 1989) and spatial-to-temporal approaches (Moscheni et al., 1998; Paragios and Triritas, 1999; Park and Lee, 1996; Salembier et al., 1995; Wang and Adelson, 1994; Won, 1998; Won and Park, 1997). The former sequentially extracts objects by iteratively determining the successive dominant motion parameters. The regions or pixels, which conform to the dominant motion parameters, are assumed to comprise one object. The other regions or pixels are regarded as undetermined ones. The process continues to estimate the subsequent dominant motion parameters for the undetermined objects. Finally, the spatial feature is used to refine the segmentation results. For an alternative approach, spatial-to-temporal, an over-segmented image is first obtained using spatial features of regions (Cortez et al., 1995; Salembier et al., 1996; Vincent and Soille, 1991), and then a region-merging procedure to identify meaningful objects is adopted using temporal information such as motion parameters. The algorithms proposed in the literatures (Moscheni et al., 1998; Paragios and Triritas, 1999; Park and Lee, 1996; Salembier et al., 1995; Wang and Adelson, 1994) usually contain three steps: (a) an initial region is generated by using the spatial characteristics, (b) temporal motion information is estimated and used for calculating a similarity measure, and (c) an object is extracted by merging similar regions according to a spatial–temporal similarity measure.

The above two approaches belong to automatic segmentation approach. However, the segmented regions are often not meaningful; e.g., a human head may be partitioned into more than two regions. Thus, in fact they cannot be regarded as VOP extraction techniques. The other common drawback of the two approaches is that many manually tuned thresholds are defined in the algorithms. The tuning of threshold values is difficult and inefficient because threshold values are always image sensitive. Furthermore, significant computation is required for both approaches.

A video object represents a meaningful entity in the world, such as a ball, a building, a human body, an aircraft, etc. (Gu and Lee, 1998). It might not be visually homogeneous in any physical sense. Video segmentation is to get a partition consisting of homogeneous spatial/temporal or spatial–temporal regions according to any given visual homogeneity criterion. Therefore, to achieve object extraction successfully, complex a priori knowledge should be included in the video segmentation. This may increase the system complexity and computational load significantly. Even so, to extract VOP accurately and reliably in an automatic manner for general video sequences is still a big challenge (Meier and Ngan, 1998). Semiautomatic methods that get some input from humans (Choi et al., 1998), e.g., by drawing initial boundary of object (Gu and Lee, 1998; Herrmann et al., 1999) were presented to alleviate the problem. However, these approaches may not be suitable for object-based video coding due to the need of user interaction.

This paper aims to develop a computationally efficient technique that can automatically extract a VOP without user interaction. To reduce the problem to a more