Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 1159–1172
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Automatic object extraction and reconstruction in active video

Ye Lu, Ze-Nian Li*

School of Computing Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6

Received 15 August 2006; received in revised form 11 May 2007; accepted 2 July 2007

Abstract

A new method of video object extraction is proposed to automatically extract the object of interest from actively acquired videos. Traditional video object extraction techniques often operate under the assumption of homogeneous object motion and extract various parts of the video that are motion consistent as objects. In contrast, the proposed active video object extraction (AVOE) approach assumes that the object of interest is being actively tracked by a non-calibrated camera under general motion and classifies the possible movements of the camera that result in the 2D motion patterns as recovered from the image sequence. Consequently, the AVOE method is able to extract the single object of interest from the active video. We formalize the AVOE process using notions from Gestalt psychology. We define a new Gestalt factor called "shift and hold" and present 2D object extraction algorithms. Moreover, since an active video sequence naturally contains multiple views of the object of interest, we demonstrate that these views can be combined to form a single 3D object regardless of whether the object is static or moving in the video.

Keywords: Active video; Camera motions; Shift and hold; Object extraction; 3D object reconstruction

1. Introduction

Fully automatic extraction of semantically meaningful objects from visual data is one of the ultimate aspirations in computer vision and pattern recognition. In addition to the obvious academic interest in this problem, there is a wide array of practical applications that can benefit tremendously from successful object extraction algorithms. One application that can immediately take advantage of object extraction is video compression. A video compression engine can selectively compress objects with higher bit-rates to produce subjectively pleasing results while lowering the bit-rates used to compress less important regions to maintain storage and transmission efficiency. Furthermore, with the proliferation of digital media, rapid searching and retrieval of multimedia data are of paramount importance to industries such as communication, education, and entertainment. It is widely believed that object extraction is the key to more efficient, accurate, and user friendly implementations of such systems. Last but not least, advanced functionalities in surveillance systems such as recognition of suspicious actions and identification of known individuals can be made much simpler with the availability of extracted objects [1].

Digital video carries rich multimedia information and it involves an insurmountable amount of data. For interactive use of the video data, the research community has been focusing on newer standards MPEG-4/H.264 and MPEG-7 where the notion of *video object* (*VO*) is the key, because in most cases VOs and their behavior are the contents! MPEG-4/H.264 has specified many VO-based coding methods. However, one thing was made clear, MPEG-4/H.264 (as other MPEG standards) is a decoding standard. The message is that we do not yet know how to accurately extract VOs.

It is observed that, in general, videos can be classified into two types: *passive video* and *active video*. A video produced by a static surveillance camera is a good example of the former. The camera's function is to (passively) record all objects passing by in front of it. Because of various security concerns, vast amount of this type of video data is generated daily and various software/systems (such as Blue Eye Video) have been developed for automated processing and analysis of these data.

* Corresponding author. Tel.: +1 604 291 3761.
E-mail addresses: yel@cs.sfu.ca (Y. Lu), li@cs.sfu.ca (Z.-N. Li).

0031-3203/$30.00
doi:10.1016/j.patcog.2007.07.015

However, video generated by an active vision system, such as our eyes, will not look like that. In general, digital videos taken by human subjects are more purposive. Typical examples will be filming, professional video cameramen covering sporting events, or an amateur shooting at a tourist scene (e.g., buildings, sculptures/statues, and activities of crowd/people). We call the video thus produced *active video*.

Active videos are very much object-centered, and often exhibit prominent catching and holding behaviors of the human operator. In order to capture the object-of-interest and its movements, it is common for the videographer to initiate various camera movements. Now the rapid pan/tilt movement is analogous to saccades, which is often triggered by object movements or distinct visual features (color, texture, shape, etc.) in the periphery, indicating a shift of attention. When dealing with moving objects, smooth (and usually not so rapid) pan/tilt movements are used for smooth pursuit. When multiple views of the object are desirable, we will witness body movement of the videographer. In professional filming, such movements are often facilitated by sliding rails and moving platforms. It should be apparent that active video is by definition object-based and full of actions.

Object extraction can be considered as a process of identifying an arbitrary collection of image regions that are usually not coherent in low level image features or motion, but somehow form a semantically meaningful entity called an "object". The lack of clear and rigorous definition of what an object is makes this problem exceptionally difficult to solve. Traditional methods of object segmentation follows the configuration laid out by Marr [2] which takes a passive approach by casting the computer as an observer from which useful information is gathered and processed. Although there have been many fruitful results along this line of research, it is very difficult to perform high level vision tasks without active participation from the vision system. It is precisely for this reason that *active vision* is proposed [3,4] for which efforts are made for computer controlled cameras to actively participate in the visual perception process, much similar to the body and eye movements of human vision systems [5]. Of course, when the camera is operated by a human being as in the case of movie making and even home video making, the lines of reasoning advocated in active vision research can essentially be reversed to form a bridge to connect the conceptual gap between the visual world and the underlying semantic meanings. From here on, we shall assume that the input image sequences or videos are acquired by intelligent active vision systems or in most cases by human beings. We thus use the term *object extraction* as opposed to the term *object segmentation* to reflect the active nature of our input data.

In this paper, we introduce a new Gestalt factor called *shift and hold* that describes the motion pattern of the potential object of interest on the image plane. We then develop the required algorithms to extract image regions corresponding to the particular motion pattern that we seek. These image regions form the object of interest and can be tracked throughout the sequence. This is our general strategy for active video object extraction or AVOE for short.

Computing 3D models from 2D views is an important but yet difficult problem in computer vision. An immediate application of visual 3D modeling through 2D views is video indexing and retrieval. If accurate 3D object models can be computed from video sequences, the retrieval system can extract useful 3D shape information from them and use this information to search for similar objects as well as eliminate false matches through shape verification. In viewing this need for 3D object models, we present our AVOE and reconstruction algorithm which extracts objects of interest from active videos and integrates various views of the same object into a single unified 3D surface model. In order to reconstruct the *Euclidean* shape of the object of interest, it is necessary to determine the calibration of the camera. However, since no calibration object was present at the time when the video was taken, traditional calibration method cannot be applied. Instead, we perform a procedure called *self-calibration* to determine the internal parameters of the camera without using any pre-made calibration objects.

2. Shift and hold: a new gestalt factor

The Gestaltist's view of perceptual organization found in 2D images provides much of the underlying principles behind modern image segmentation algorithms. Although these organizational principles can be applied in a similar manner to image sequences (or videos) to perform figure and ground segregation, doing so will likely fail to exploit the richness of information contained within image sequences and may not capture the intentions of the author of the video. In this section, we introduce a new Gestalt factor called *shift and hold* which bridges the gap between static images and video sequences.

2.1. Motivation

Figure and ground segregation is not only an interesting problem in the academic sense but also has a large number of potential practical applications. Gestalt psychology defines a number of factors that can aid in figure and ground segregation on static 2D images [6]. However, because 2D images are perspective projections of the 3D world, much information is lost during the projection process. As a result, it is sometimes extremely ambiguous to separate the figure from the ground even after we apply these Gestalt principles. Some examples of well-known ambiguities are shown in Fig. 1. These ambiguities occur when both the black and white regions have valid semantic interpretations. It is evident that these ambiguities remain even to the human eyes. The fact that our biological vision system rarely produces ambiguous interpretations of the world suggests that most of these artificially designed 2D visual ambiguities can be resolved when we attempt to perceive objects in 3D using various cues such as lighting, shading, shadows, and through the *stereopsis* process.

Fig. 2 shows another illustration of Rubin's vase. In that illustration, there are various visual cues on the vase so that it is immediately perceived as the figure while the black areas are the ground. Comparing to Fig. 1(a) where figure and ground reversals often occur, the shadings, reflections, the deformations