Computer Vision and Image Understanding 141 (2015) 52–66

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Object motion analysis description in stereo video content ☆

Theodoris Theodoridis, Konstantinos Papachristou*, Nikos Nikolaidis, Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Box 451, 54124 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 19 December 2014
Accepted 5 July 2015
Available online 11 July 2015

Keywords:
Motion analysis
Motion characterization
Stereo video
Semantic labeling

ABSTRACT

The efficient search and retrieval of the increasing volume of stereo videos drives the need for the semantic description of its content. The analysis and description of the disparity (depth) data available on such videos, offers extra information, either for developing better video content search algorithms, or for improving the 3D viewing experience. Taking the above into account, the purpose of this paper is twofold. First, to provide a mathematical analysis of the relation of object motion between world and display space and on how disparity changes affect the 3D viewing experience. Second, to propose algorithms for semantically characterizing the motion of an object or object ensembles along any of the X, Y, Z axis. Experimental results of the proposed algorithms for semantic motion description in stereo video content are given.

1. Introduction

In recent years, the production of 3D movies and 3D video has been growing significantly. A large number of 3D movies have been released and some of them, e.g. Avatar [1] had great success. These box-office successes have boosted (a) the delivery of 3D productions, such as movies and documentaries, to home or to cinema theaters through 3D display technologies [2] and (b) the 3DTV broadcasting of various events, such as sports [3], [4], for a high quality 3D viewing experience. Furthermore, virtual reality systems for computer graphics, entertainment and education, which use stereo video technology, have been developed [5–7]. 3D video devices such as laptops, cameras, mobile phones, TV, projectors are now widely available for professional and non-professional users [1]. Because of the 3D movie success, several tools have been developed for the production and editing of 3D content [8,9].

Since 3DTV content is now widely available, it must be semantically described toward fast 3D video content search and retrieval. Analysis of stereoscopic video has the advantage of deriving information that cannot be inferred from single-view video, such as 3D object position through depth/disparity information. Depth information can also be obtained from multiple synchronized video streams [10–12]. MPEG-4 offers a set of motion descriptors for the representation of motion of a trajectory [13]. 3D motion descriptors include the world coordinates and time information. In this paper, we propose the adoption such 3D descriptors for the extraction semantic labels such as “an object approaches the camera” or “two objects approach each other”. Such semantic description is only possible using 3D descriptors instead of 2D descriptors. In this paper, we concentrate on 3D object motion description in stereo video content. Various algorithms for semantic labeling of human, object or object ensemble motion are proposed. We utilize the depth information, which is implicitly available through disparity estimation between the left and right views, to examine various cases, where camera calibration information and/or viewing parameters may or may not be available, assuming that there are no camera motion and fixed intrinsic parameters. For example, we can characterize video segments, where an object approaches the camera or where two objects approach each other in the real world. It should be noted that the proposed algorithms can be applied in the case of a calibrated Kinect camera as well [14]. Indeed, a lot of works investigate the 3D reconstruction of object trajectories [15–17]. The novelty of the proposed algorithms is the object motion analysis providing semantic labels. Such semantic stereo video content description is very useful in various applications, varying from video surveillance and 3D video annotations archiving, indexing and retrieval to implementation of better audiovisual editing tools and intelligent content manipulation. Such characterization is not possible in classical single view video, without knowing depth information to get 3D position/motion clues [8]. Furthermore, such characterizations can be used for detecting various stereo quality effects [8]. For example, if an object having strong negative disparity has been labeled as moving along the x axis toward the left/right image border, then it is likely that a left/right stereo window violation may arise. The distance between foreground objects and the background influences the entire amount of depth information (depth budget) of the scene during display.

☆ This paper has been recommended for acceptance by Y. Aloimonos.

* Corresponding author.
E-mail addresses: kpapaxristou@aiia.csd.auth.gr, konpapac23@hotmail.com (K. Papachristou), nikolaid@aiia.csd.auth.gr (N. Nikolaidis), pitas@aiia.csd.auth.gr (I. Pitas).

http://dx.doi.org/10.1016/j.cviu.2015.07.002

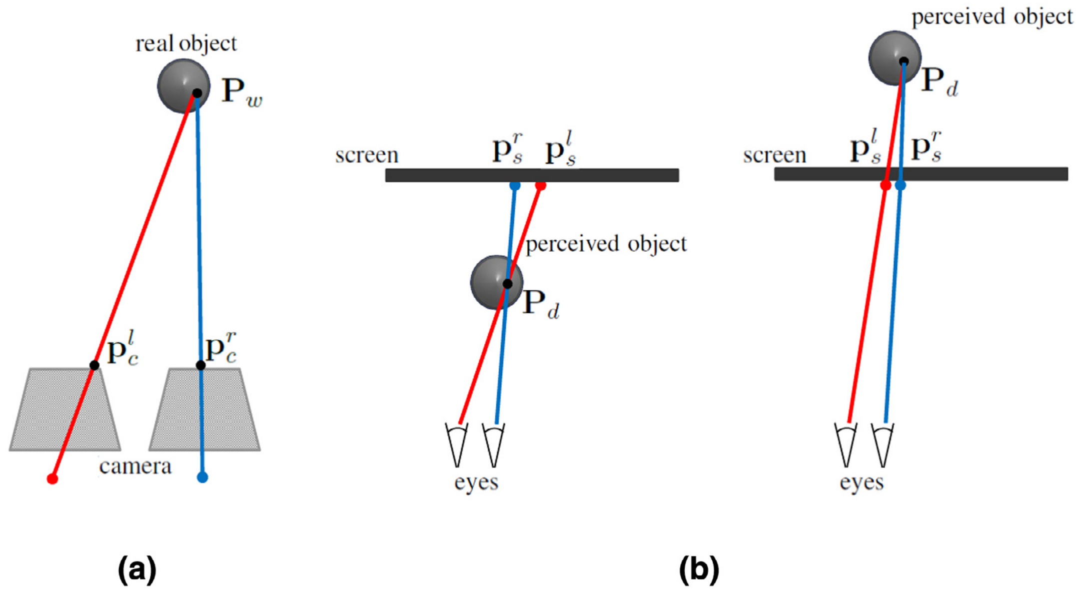

Fig. 1. (a) Stereo video capture and (b) display.

Furthermore, we examine how the viewer perceives object motion during stereo display. Typically, stereo video is shot with a stereo camera to display objects residing and moving in the world space (X_w, Y_w, Z_w). The acquired stereo video depends on the stereo camera parameters, e.g., focal length and the baseline distance [8]. When displayed, the perceived object position and motion occurs in the display (theater) space (X_d, Y_d, Z_d). The perceived video content depends on the viewing parameters, e.g., the screen size and the viewing distance. The real and the perceived object motion may differ, depending on the camera and viewing parameters, as well as on stereo content manipulations [8]. Specifically, we assume that an object is moving with a known motion type (e.g., constant speed motion along the Z_w axis) and we determine what motion is perceived by the viewer. We examine various simple motion types, such as motion with constant velocity or constant acceleration along axes X_w, Y_w or Z_w. This analysis is very useful for avoiding cases where excessive motion particularly along the Z_w axis can cause viewing discomfort [18]. In addition, we elaborate on how disparity modifications affect the perceived position of the object in the theater space with respect to the viewer. This is very important in the stereo video post-production, when the scene depth is adapted for visually stressing important scenes or for ensuring visual comfort [8]. In this respect, the relationship between the viewer's angular eye velocity and object motion in the world space is very important.

The main novel contributions of this paper are:

1. we study (Section 3) object motion in stereo video content by providing a novel mathematical analysis. The object position, velocity and acceleration are examined in various simple motion types. In addition, we study the relationship between the viewer's angular eye velocity and object motion, in order to examine how the viewer perceives object motion during stereo display. In the same theoretical context, we elaborate on how disparity modifications affect the perceived position of the object in the theater space.
2. We provide (Section 4) novel algorithms for the semantic description/characterization of object motion in stereo video content along the horizontal, vertical and depth axis, as well as characterizations of relative motion of pairs of objects (whether the objects approach each other or move away).

These two contributions (theoretical, algorithmic) refer to different motion characteristics and thus are not related.

The paper extends the work in [19] and [20] by including (a) the study of object motion in stereo video content providing a novel mathematical analysis and (b) the assessment of the robustness of the presented motion labeling methods in challenging scenes recorded outdoors in realistic conditions.

The rest of the paper is organized as follows. In Section 2, the geometry of the stereo camera and of the display system is discussed. The transformations between the different coordinate systems of the world, stereo camera, screen and display (theater) space are given for two stereo camera setups, the parallel and converging ones. Section 3 contains the mathematical analysis for the relation between world and display system, the impact of screen disparity modifications on object position during display and the relation between object and viewer's eye motion. In Section 4, algorithms for characterizing object and object ensemble motion are proposed. In Section 5, experimental results for motion characterization are presented. Finally, concluding remarks are given in Section 6.

2. Stereo video acquisition and display geometry

In stereo video, a 3D scene is captured by a stereo camera (a video camera pair), as shown in Fig. 1(a). A point of interest $\mathbf{P}_w = [X_w, Y_w, Z_w]^\top$ in the 3D world space is projected on the left and right image plane positions $\mathbf{p}_c^l = \left[x_c^l, y_c^l\right]^\top$ and $\mathbf{p}_c^r = [x_c^r, y_c^r]^\top$, respectively. For stereo video display, both images are projected (mapped) on the display screen plane locations $\mathbf{p}_s^l = \left[x_s^l, y_s^l\right]^\top$ and $\mathbf{p}_s^r = [x_s^r, y_s^r]^\top$, respectively, as shown in Fig. 1(b). During display, the point $\mathbf{P}_d = [X_d, Y_d, Z_d]^\top$ which corresponds to $\mathbf{P}_w$ is perceived by the viewer in front of, on or behind the screen in the display (theater) space, as shown in Fig. 1(b), if the disparity $d = x_s^r - x_s^l$ is negative or positive, respectively.

In this section, we describe in more detail the geometrical relations between the world and theater space coordinates for two types of stereo camera setups, the parallel [21], which is the most common case, and the converging one [22].

2.1. Parallel stereo camera setup

The geometry of a stereo camera with parallel optical axes is shown in Fig. 2. The centers of projection and the projection planes of the left and right camera are denoted by the points $\mathbf{O}_l$, $\mathbf{O}_r$ and $\mathcal{T}_l$,