Object Motion Analysis Description In Stereo Video Content

T. Theodoridis[a], K. Papachristou[a], N. Nikolaidis[a], I. Pitas[a]

[a] *Aristotle University of Thessaloniki, Department of Informatics*
Box 451, 54124 Thessaloniki, Greece

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.

Keywords: Motion analysis, motion characterization, stereo video, semantic labelling.

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 high a quality 3D viewing experience. Furthermore, virtual

Email addresses: nikolaid@aiia.csd.auth.gr (N. Nikolaidis), pitas@aiia.csd.auth.gr (I. Pitas)

reality systems for computer graphics, entertainment and education, which use stereo video technology, have been developed [5], [6], [7]. 3D video devices such us 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 towards 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, 11, 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 labelling 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 an calibrated Kinect camera as well [14]. Indeed, a lot of works investigate the 3D reconstruction of object trajectories [15, 16, 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 labelled as moving along the x axis towards 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.

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 viewers 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.

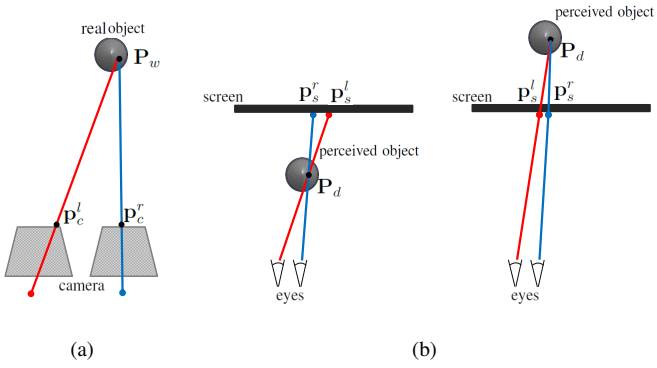
The paper extends the work in [19] and [20] by including (a) the study of object motion in stereo video content providing a novel mathematical analysis and (b) the assessment of the robustness of the presented motion labelling methods in challenging scenes recorded outdoors in realistic conditions.

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 Figure1(a). A point of interest $\mathbf{P}_w = [X_w, Y_w, Z_w]^\intercal$ in the 3D world space is projected on the left and right image plane positions $\mathbf{p}_c^l = \left[x_c^l, y_c^l\right]^\intercal$ and $\mathbf{p}_c^r = [x_c^r, y_c^r]^\intercal$, respectively. For stereo video display, both images are projected (mapped) on the display screen plane locations $\mathbf{p}_s^l = \left[x_s^l, y_s^l\right]^\intercal$ and $\mathbf{p}_s^r = [x_s^r, y_s^r]^\intercal$, respectively, as

shown in Figure1(b). During display, the point $\mathbf{P}_d = [X_d, Y_d, Z_d]^\intercal$ 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 Figure1(b), if the disparity $d = x_s^r - x_s^l$ is negative or positive, respectively.



(a) (b)

Figure 1: a) stereo video capture and b) display.

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 Figure 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 $\mathcal{T}_l$, $\mathcal{T}_r$, respectively. The distances between the two camera centers and between the camera center of projection and the projection plane are the baseline distance T_c and the camera focal length f. The midpoint $\mathbf{O}_c$ of the baseline is the center of the world coordinate system (X_w, Y_w, Z_w). The world coordinate axis X_w can be transformed into the left/right camera axes X_w^l, X_w^r by a translation by $\pm T_c/2$. A point of interest $\mathbf{P}_w = [X_w, Y_w, Z_w]^\intercal$ in the world space is projected on the left and right image planes at the points $\mathbf{p}_c^l = \left[x_c^l, y_c^l\right]^\intercal$ and $\mathbf{p}_c^r = [x_c^r, y_c^r]^\intercal$ respectively, while the points $\mathbf{P}_w^l = \left[X_w^l, Y_w^l, Z_w^l\right]^\intercal$ and $\mathbf{P}_w^r = [X_w^r, Y_w^r, Z_w^r]^\intercal$ refer to the same point $\mathbf{P}_w$ with respect to the left and right camera coordinate systems, respectively. The projections $\mathbf{p}_c^l$ and $\mathbf{p}_c^r$ are related with the 3D points $\mathbf{P}_w^l$ and $\mathbf{P}_w^r$ using

perspective projection [21]:

$$x_c^l = f\frac{X_w^l}{Z_w^l}, \quad y_c^l = f\frac{Y_w^l}{Z_w^l}, \quad x_c^r = f\frac{X_w^r}{Z_w^r}, \quad y_c^r = f\frac{Y_w^r}{Z_w^r}. \tag{1}$$

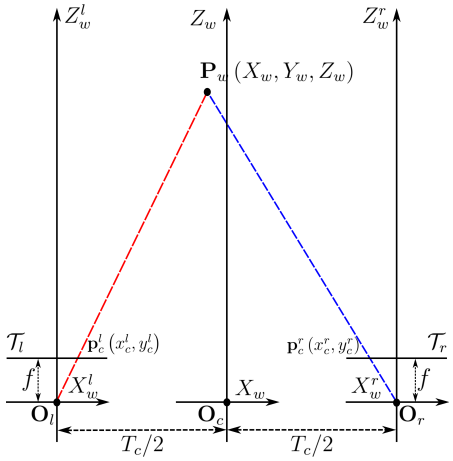


Figure 2: Parallel stereo camera geometry.

Thus, the following equations give us the transform from the world space to the camera system coordinates:

$$x_c^l = f\frac{X_w + \frac{T_c}{2}}{Z_w}, \quad y_c^l = f\frac{Y_w}{Z_w}, \quad x_c^r = f\frac{X_w - \frac{T_c}{2}}{Z_w}, \quad y_c^r = f\frac{Y_w}{Z_w}. \tag{2}$$

It is well known that the $\mathbf{P}_w$ world space coordinates can be recovered from the $\mathbf{p}_c^l, \mathbf{p}_c^r$ projections, as follows [21]:

$$Z_w = -\frac{fT_c}{d_c}, \quad X_w = -\frac{T_c(x_c^l + x_c^r)}{2d_c}, \quad Y_w = -\frac{T_c y_c^l}{d_c} = -\frac{T_c y_c^r}{d_c}, \tag{3}$$

where $d_c = x_c^r - x_c^l$ is the stereo disparity. In the case of the parallel camera setup, we always have negative disparities:

$$d_c = -f\frac{T_c}{Z_w} < 0. \tag{4}$$

The geometry of the display (theater) space is shown in Figure 3. T_e is the distance between the left/right eyes (typically, 60 mm) [23]. The distance from the viewer's eye pupil centers ($\mathbf{e}_l$ and $\mathbf{e}_r$, respectively) to the screen is denoted by T_d. The origin $\mathbf{O}_d$

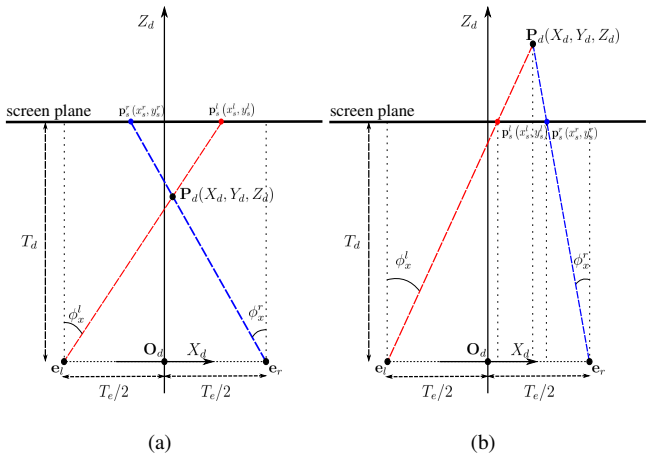


Figure 3: Stereo display system geometry for a) negative and b) positive screen disparity.

of the display coordinate system (X_d, Y_d, Z_d) is placed at the midpoint between the eyes. The X_d axis is parallel to the eye baseline. The Z_d, Y_d axes are perpendicular to the screen and $X_d Z_d$ planes, respectively. During stereo image display, the mapping of the projections $\mathbf{p}_c^l$ and $\mathbf{p}_c^r$ to the screen plane $\mathbf{p}_s^l = \left[x_s^l, y_s^l\right]^\intercal$ and $\mathbf{p}_s^r = [x_s^r, y_s^r]^\intercal$ is achieved by scaling using a factor $m = w_s/w_c$, where w_s is the width of the screen and w_c the width of the camera sensor,

$$x_s^l = m x_c^l, \quad y_s^l = m y_c^l, \quad x_s^r = m x_c^r, \quad y_s^r = m y_c^r, \tag{5}$$

that magnifies the image, according to the screen size, while the screen center coordinate (x_s, y_s) coincides with the shifted by T_c left/right image plane coordinate $\left(x_c^l, y_c^l\right)$, (x_c^r, y_c^r) centers, so that they coincide. Here, the distance of x_s^l and x_s^r, $d_s = x_s^r - x_s^l$, is the *screen disparity*. The resulting perceived object position is in front of, on and behind the screen for negative, zero and positive screen disparity, respectively, as shown in Figure 3a,b. The perceived location $\mathbf{P}_d(X_d, Y_d, Z_d)$ of the point $\mathbf{P}_w$ can be found using triangle $(\mathbf{p}_s^l \mathbf{P}_d \mathbf{p}_s^r)$, $(\mathbf{e}_l \mathbf{P}_d \mathbf{e}_r)$ similarities [22]:

$$Z_d = \frac{T_d T_e}{T_e - d_s}, \tag{6}$$

$$X_d = \frac{T_e(x_s^l + x_s^r)}{2(T_e - d_s)}, \quad Y_d = \frac{T_e(y_s^l + y_s^r)}{2(T_e - d_s)}. \tag{7}$$

Since in the parallel camera setup we always have negative disparities d_c and thus $T_e - d_s > T_e$, all objects appear in front of the screen $Z_d < T_d$. It can be easily proven that the coordinate transformation from the camera image plane to display space is given by:

$$X_d = \frac{mT_e(x_c^l + x_c^r)}{2(T_e - md_c)}, \quad Y_d = \frac{mT_e(y_c^l + y_c^r)}{2(T_e - md_c)}, \quad Z_d = \frac{T_d T_e}{T_e - md_c}. \tag{8}$$

Finally, we can compute the overall coordinate transformation from world space to display space

$$X_d = \frac{mfT_e X_w}{mfT_c + T_e Z_w}, Y_d = \frac{mfT_e Y_w}{mfT_c + T_e Z_w}, Z_d = \frac{T_d T_e Z_w}{mfT_c + T_e Z_w}. \tag{9}$$

The display geometry shown in Figure 3 describes well stereo projection in theater, TV, computer and mobile phone screens, but not in virtual reality systems (head-mounted displays) [24].

2.2. Converging Stereo Camera Setup

In this case, the optical axes of the left and right camera form an angle θ with the coordinate axis Z_w, as shown in Figure 4. The origin $\mathbf{O}_c$ of the world space coordinate system is placed at the midpoint between the left and right camera centers. The two camera axes converge on the point $\mathbf{O}_z$ at distance T_z along the Z_w axis. A point of interest $\mathbf{P}_w = [X_w, Y_w, Z_w]^\intercal$ in the world space, which is projected on the left and right image planes at the points $\mathbf{p}_c^l = \left[x_c^l, y_c^l\right]^\intercal$ and $\mathbf{p}_c^r = [x_c^r, y_c^r]^\intercal$, respectively, can be transformed into the left or right camera system by a translation by $T_c/2$ or $-T_c/2$, respectively, followed by a rotation by angle $-\theta$ or θ about the Y_w axis, respectively:

$$\begin{bmatrix} X_w^l \\ Y_w^l \\ Z_w^l \end{bmatrix} = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} X_w + \frac{T_c}{2} \\ Y_w \\ Z_w \end{bmatrix}, \tag{10}$$

$$\begin{bmatrix} X_w^r \\ Y_w^r \\ Z_w^r \end{bmatrix} = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} X_w - \frac{T_c}{2} \\ Y_w \\ Z_w \end{bmatrix}. \tag{11}$$

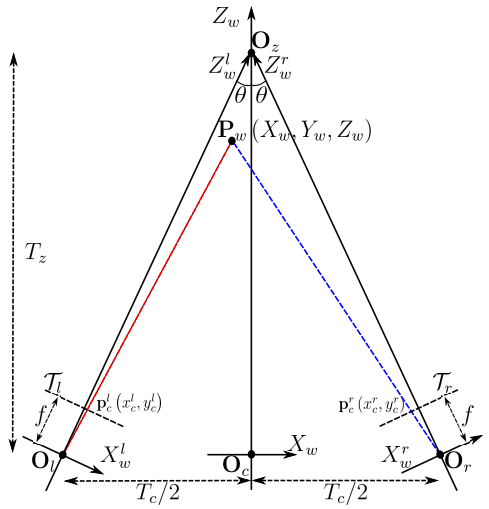


Figure 4: Converging stereo camera setup geometry.

Using (1), the following equations transform the world space coordinates to the left/right camera coordinates:

$$
\begin{aligned}
x_c^l &= f\frac{\left(X_w + \frac{T_c}{2}\right)\cos\theta - Z_w\sin\theta}{\left(X_w + \frac{T_c}{2}\right)\sin\theta + Z_w\cos\theta} \\
&= f\tan\left(\arctan\left(\frac{X_w + \frac{T_c}{2}}{Z_w}\right) - \theta\right) \qquad (12)\\
y_c^l &= f\frac{Y_w}{\left(X_w + \frac{T_c}{2}\right)\sin\theta + Z_w\cos\theta}, \qquad (13)\\
x_c^r &= f\frac{\left(X_w - \frac{T_c}{2}\right)\cos\theta + Z_w\sin\theta}{-\left(X_w - \frac{T_c}{2}\right)\sin\theta + Z_w\cos\theta} \\
&= -f\tan\left(\arctan\left(\frac{-X_w + \frac{T_c}{2}}{Z_w}\right) - \theta\right), \qquad (14)\\
y_c^r &= f\frac{Y_w}{-\left(X_w - \frac{T_c}{2}\right)\sin\theta + Z_w\cos\theta}. \qquad (15)
\end{aligned}
$$

For very small angles θ (12)-(15) can be simplified using $\cos\theta \simeq 1$, $\sin\theta \simeq \theta$ rad. When $\theta = 0$, then equations (12)-(15) collapse to (8)-(9). As proven in the Appendix A, the following equations can be used, in order to revert from the left/right camera

coordinates into the world space coordinates:

$$X_w = T_c \frac{x_c^l + \tan\theta\left(f + \frac{x_c^l x_c^r}{f} + x_c^r \tan\theta\right)}{x_c^l - x_c^r + \tan\theta\left(2f + 2\frac{x_c^l x_c^r}{f} - x_c^l \tan\theta + x_c^r \tan\theta\right)} - \frac{T_c}{2}, \tag{16}$$

$$Y_w = T_c \frac{y_c^l}{f} \frac{\cos\left(\arctan\left(\frac{x_c^l}{f}\right) + \theta\right)\cos\left(\arctan\left(\frac{x_c^l}{f}\right)\right)}{\sin\left(\arctan\left(\frac{x_c^l}{f}\right) + \arctan\left(\frac{x_c^r}{f}\right) + 2\theta\right)}, \tag{17}$$

$$Z_w = T_c \frac{f - \left(x_c^l - x_c^r + \frac{x_c^l x_c^r}{f}\tan\theta\right)\tan\theta}{x_c^l - x_c^r + \tan\theta\left(2f + 2\frac{x_c^l x_c^r}{f} - x_c^l \tan\theta + x_c^r \tan\theta\right)}. \tag{18}$$

Following the same methodology as in the parallel setup, the transformations from camera plane to the 3D display space are given by (5), (6) and (7), respectively. For the case of $X_w = 0$, it can easily be proven that, when $Z_w > T_z$, the object appears behind the screen ($Z_d > T_d$), while for $Z_w < T_z$, the object appears in front of the screen, as exemplified in Figure 3a. This is the primary reason for using the converging camera setup in 3D cinematography. However, only smalls θs are used, because otherwise the so-called keystroke effect is very visible [8].

Finally, the overall coordinate transformation from world space to display space is given [22] by the equations (19)-(21).

$$X_d = \frac{mfT_e\left(\tan\left(\arctan\left(\frac{X_w + \frac{T_c}{2}}{Z_w}\right) - \theta\right) - \tan\left(\arctan\left(\frac{-X_w + \frac{T_c}{2}}{Z_w}\right) - \theta\right)\right)}{2T_e + 2mf\left(\tan\left(\arctan\left(\frac{-X_w + \frac{T_c}{2}}{Z_w}\right) - \theta\right) + \tan\left(\arctan\left(\frac{X_w + \frac{T_c}{2}}{Z_w}\right) - \theta\right)\right)} \tag{19}$$

$$Y_d = \frac{mT_e\left(f\frac{Y_w}{\left(X_w + \frac{T_c}{2}\right)\sin\theta + Z_w\cos\theta} + f\frac{Y_w}{-\left(X_w - \frac{T_c}{2}\right)\sin\theta + Z_w\cos\theta}\right)}{2T_e + 2mf\left(\tan\left(\arctan\left(\frac{-X_w + \frac{T_c}{2}}{Z_w}\right) - \theta\right) + \tan\left(\arctan\left(\frac{X_w + \frac{T_c}{2}}{Z_w}\right) - \theta\right)\right)} \tag{20}$$

$$Z_d = \frac{T_d T_e}{T_e + mf\left(\tan\left(\arctan\left(\frac{-X_w + \frac{T_c}{2}}{Z_w}\right) - \theta\right) + \tan\left(\arctan\left(\frac{X_w + \frac{T_c}{2}}{Z_w}\right) - \theta\right)\right)} \tag{21}$$

When $\theta = 0$, (16) - (18) and (19) - (21) collapse to the parallel setup equations (3) and (9).

3. Mathematical Object Motion Analysis

In this section, the 3D object motion in stereo vision is mathematically treated. No such treatment exists in the literature, at least to the authors' knowledge. In subsection 3.1, we examine the true 3D object motion compared to the perceived 3D motion of the displayed object in the display space. In subsection 3.2, we elaborate on how the change of screen projections affects stereo video content display. Finally, the effect of the perceived object motion on visual comfort is presented in subsection 3.3.

3.1. Motion mapping between World and Display Space

In this section, we analyse the perceived object motion during stereo video acquisition and display, assuming that the object motion trajectory in world space $[X_w(t), Y_w(t), Z_w(t)]^\intercal$ is known. We consider the parallel camera setup geometry. The perceived motion speed and acceleration can be derived by differentiating (9):

$$v_{Z_d}(t) = \frac{T_e T_d T_c f m Z'_w(t)}{(mfT_c + T_e Z_w(t))^2}, \tag{22}$$

$$a_{Z_d}(t) = -\frac{T_e T_d T_c f m \left(-2T_e Z'_w(t)^2 + (T_c mf + TeZ_w(t)) Z''_w(t)\right)}{(mfT_c + T_e Z_w(t))^3}, \tag{23}$$

$$v_{X_d}(t) = \frac{mfT_e \left((mfT_c + T_e Z_w(t)) X'_w(t) - T_e X_w(t) Z'_w(t)\right)}{(mfT_c + T_e Z_w(t))^2}, \tag{24}$$

$$\begin{aligned} a_{X_d}(t) = {} & \frac{mfT_e \left((mfT_c + T_e Z_w(t)) X''_w(t) - T_e X_w(t) Z''_w(t)\right)}{(mfT_c + T_e Z_w(t))^2} \\ & - \frac{2mfT_e^2 Z'_w(t) \left((mfT_c + T_e Z_w(t)) X'_w(t) - T_e X_w(t) Z'_w(t)\right)}{(mfT_c + T_e Z_w(t))^3}. \end{aligned} \tag{25}$$

Similar equations can be derived for motion speed and acceleration along the Y_d axis. The following two cases are of special interest:

a) If the object is moving along the Z_w world axis with constant velocity $Z_w(t) = Z_{w_0} + v_{Z_w} t$, its perceived motion along the Z_d axis has no constant velocity anymore:

$$Z_d(t) = \frac{T_e T_d \left(Z_{w_0} + v_{Z_w} t\right)}{mfT_c + T_e \left(Z_{w_0} + v_{Z_w} t\right)}, \tag{26}$$

$$v_{Z_d}(t) = \frac{T_e T_d T_c f m v_{Z_w}}{\left(mfT_c + T_e \left(Z_{w_0} + v_{Z_w} t\right)\right)^2}, \tag{27}$$

$$a_{Z_d}(t) = -\frac{2T_c T_e^2 T_d f m v_{Z_w}^2}{\left(mfT_c + T_e \left(Z_{w_0} + v_{Z_w} t\right)\right)^3}. \tag{28}$$

b) If the object is moving along the Z_w world axis with constant acceleration $Z_w(t) = Z_{w_0} + \frac{1}{2} a_{Z_w} t^2$, the perceived motion along the Z_d axis is even more complicated:

$$Z_d(t) = \frac{T_e T_d \left(a_{Z_w} t^2 + 2Z_{w_0}\right)}{2mfT_c + T_e \left(a_{Z_w} t^2 + 2Z_{w_0}\right)}, \tag{29}$$

$$v_{Z_d}(t) = \frac{4T_e T_d m f T_c a_{Z_w} t}{\left(2mfT_c + T_e \left(a_{Z_w} t^2 + 2Z_{w_0}\right)\right)^2}, \tag{30}$$

$$a_{Z_d}(t) = -\frac{mfT_e T_d T_c \left(12T_e a_{Z_w} t^2 - 8mfT_c - 8T_e Z_{w_0}\right) a_{Z_w}}{\left(2mfT_c + T_e \left(a_{Z_w} t^2 + 2Z_{w_0}\right)\right)^3}. \tag{31}$$

In both cases the perceived velocity and acceleration are not constant. Additionally, under certain conditions an accelerating object may be perceived as a decelerating one.

If the object is moving along the X_w world axis with constant velocity $X_w(t) = X_{w_0} + v_{X_w} t$ and is stationary along the Z_w world axis $Z_w(t) = Z_{w_0}$, the perceived motion along axis the X_d axis has constant velocity:

$$X_d(t) = \frac{mfT_e}{mfT_c + T_e Z_{w_0}} \left(X_{w_0} + v_{X_w} t\right), \tag{32}$$

$$v_{X_d}(t) = \frac{mfT_e}{mfT_c + T_e Z_{w_0}} v_{X_w}, \tag{33}$$

$$a_{X_d}(t) = 0. \tag{34}$$

If the object is moving along the X_w world axis with constant acceleration $X_w(t) = X_{w_0} + \frac{1}{2} a_{X_w} t^2$ and is stationary along the Z_w world axis, $Z_w(t) = Z_{w_0}$, the same motion pattern applies to the perceived motion in the theater space:

$$X_d(t) = \frac{mfT_e}{mfT_c + T_e Z_{w_0}} \left(X_{w_0} + \frac{1}{2} a_{X_w} t^2 \right), \quad (35)$$

$$v_{X_d}(t) = \frac{mfT_e}{mfT_c + T_e Z_{w_0}} a_{X_w} t, \quad (36)$$

$$a_{X_d}(t) = \frac{mfT_e}{mfT_c + T_e Z_{w_0}} a_{X_w}. \quad (37)$$

In both cases the perceived velocity and acceleration are the actual world ones, scaled by a constant factor. If the object is moving along the X_w and Z_w world axes with constant velocities $X_w(t) = X_{w_0} + v_{X_w} t$, $Z_w(t) = Z_{w_0} + v_{Z_w} t$, the perceived motion pattern is very complicated.

$$X_d(t) = \frac{mfT_e}{mfT_c + T_e (Z_{w_0} + v_{Z_w} t)} (X_{w_0} + v_{X_w} t), \quad (38)$$

$$v_{X_d}(t) = \frac{mfT_e (mfT_c v_{X_w} - T_e v_{Z_w} X_{w_0} + T_e v_{X_w} Z_{w_0})}{(mfT_c + T_e (Z_{w_0} + v_{Z_w} t))^2}, \quad (39)$$

$$a_{X_d}(t) = -\frac{2mfT_e^2 v_{X_w} (mfT_c v_{X_w} - T_e v_{Z_w} X_{w_0} + T_e v_{X_w} Z_{w_0})}{(mfT_c + T_e (Z_{w_0} + v_{Z_w} t))^3}, \quad (40)$$

The case of motion along the Y_w world axis is similar to the one along the X_w axis. For the case of constant velocities along both the X_w and Z_w world axes, it is apparent that $\frac{v_{X_w}}{v_{X_d}} \neq \frac{v_{Z_w}}{v_{Z_d}}$. Thus the perceived moving object trajectory is different than the respective linear trajectory in the world space. It is clearly seen that special care should be taken when trying to display 3D moving objects, especially when the motion along the Z_w is quite irregular.

3.2. The Effects of Screen Disparity Manipulations

Let us assume that the position of the projections $\mathbf{p}_s^l = \left[x_s^l, y_s^l\right]^\intercal$ and $\mathbf{p}_s^r = [x_s^r, y_s^r]^\intercal$ of a point $\mathbf{P}_w$ on the screen can move with constant velocity. Assuming that there is no vertical disparity, we examine only x coordinates change at constant velocities u_{xl}, u_{xr}:

$$x_s^l(t) = x_{s^0}^l + v_{xl} t, \quad (41)$$

$$x_s^r(t) = x_{s^0}^r + v_{xr} t, \quad (42)$$

where $x^l_{s^0}$ and $x^r_{s^0}$ are the initial object positions on the screen plane and v_{xl} and v_{xr} indicate the corresponding velocities, having left and right direction respectively. Correspondingly, the screen disparity changes:

$$d_s(t) = x^r_{s^0} - x^l_{s^0} + (v_{xr} - v_{xl})t. \tag{43}$$

Based on the equations (6) and (7), which compute the X_d, Y_d and Z_d coordinates of $\mathbf{P}_d$ during display with respect to screen coordinates, the following equations give the $\mathbf{P}_d$ position and velocity:

$$\begin{aligned}
Z_d(t) &= \frac{T_d T_e}{T_e - d_s(0) - (v_{xr} - v_{xl})\, t}, && (44)\\
\frac{dZ_d(t)}{dt} &= \frac{T_d T_e\, (v_{xr} - v_{xl})}{(T_e - d_s(0) - (v_{xr} - v_{xl})\, t)^2}, && (45)\\
Y_d(t) &= \frac{T_e(y^l_s + y^r_s)}{2(T_e - d_s(0) - (v_{xr} - v_{xl})t)}, && (46)\\
\frac{dY_d(t)}{dt} &= \frac{T_e(y^l_s + y^r_s)\, (v_{xr} - v_{xl})}{2\,(T_e - d_s(0) - (v_{xr} - v_{xl})\, t)^2}, && (47)\\
X_d(t) &= \frac{T_e(x^r_{s^0} + v_{xr}t + x^l_{s^0} + v_{xl}t)}{2\,(T_e - d_s(0) - (v_{xr} - v_{xl})\, t)}, && (48)\\
\frac{dX_d(t)}{dt} &= \frac{T_e^2\, (v_{xr} + v_{xl}) + 2T_e\left(v_{xr}x^l_{s^0} - v_{xl}x^r_{s^0}\right)}{2\,(T_e - d_s(0) - (v_{xr} - v_{xl})\, t)^2}. && (49)
\end{aligned}$$

As expected, according to the (45) the object appears moving away from the viewer, when $v_{xr} > v_{xl}$, and approaching the viewer, when $v_{xr} < v_{xl}$. In the case of $v_{xr} = v_{xl}$, the value of Z_d does not change. Similarly, though the vertical disparity is zero, according to (47), the object appears moving downwards/upwards, when v_{xr} is bigger/smaller than v_{xl}, respectively, while in case of $v_{xr} = v_{xl}$, the value of Y_d does not change. Finally, according to (49), the cases where X_d increases, decreases and does not change are illustrated in the Figure 5.

Therefore, disparity manipulations (e.g., increase/decrease) during post-production can create significant changes in the perceived object position and motion in the display space. These effects should be better understood, in order to perform effective 3D movie post-production. It should be noted that viewing experience is also affected by motion cues and the display settings [25].

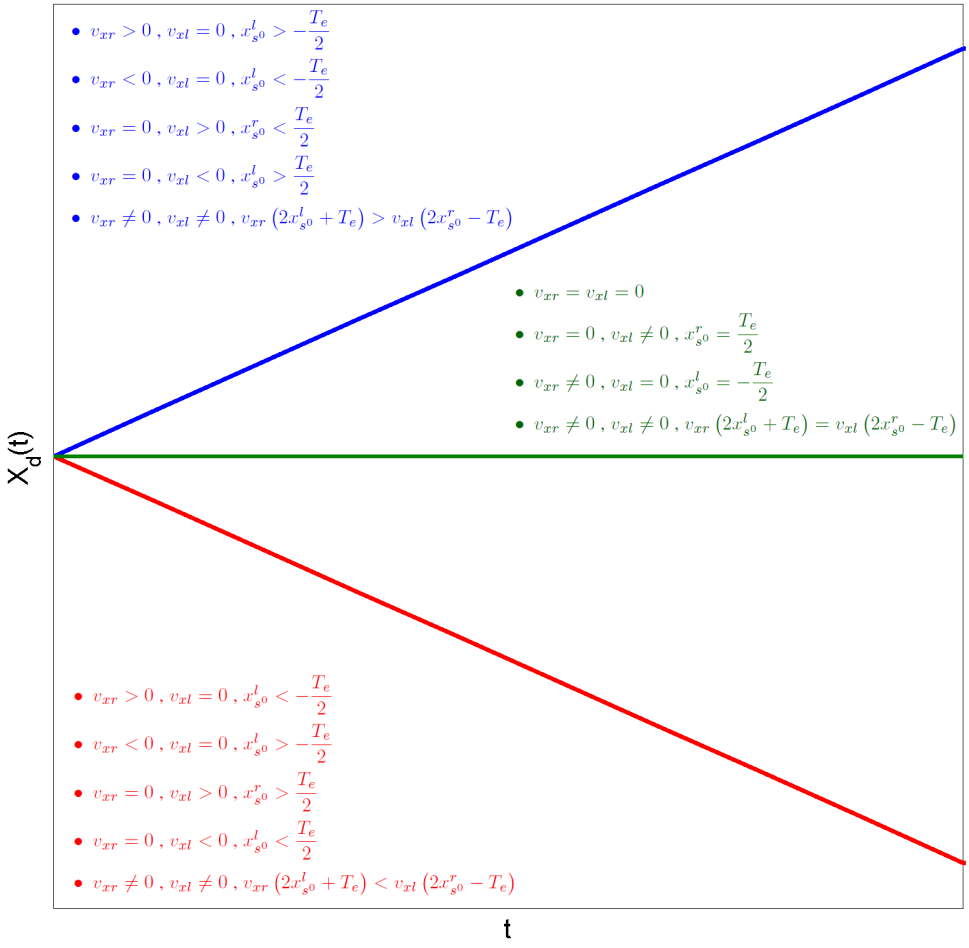


Figure 5: The cases where X_d increases, decreases and does not change.

3.3. Angular Eye Motion

When eyes view a point on the screen, they converge to the position dictated by its disparity, as shown in Figure 3. The eye convergence angles ϕ^l_x , ϕ^r_x are given by the

following equations:

$$\phi_x^l = \arctan\left(\frac{x_s^l + \frac{T_e}{2}}{T_d}\right), \tag{50}$$

$$\phi_x^r = \arctan\left(\frac{x_s^r - \frac{T_e}{2}}{T_d}\right). \tag{51}$$

The angle ϕ_y formed between the eye axis and the horizontal plane is given by:

$$\phi_y = \arctan\left(\frac{y_s^l}{T_d}\right) = \arctan\left(\frac{y_s^r}{T_d}\right). \tag{52}$$

If the camera parameters are unknown, the angular eye velocities can be derived by differentiating (50), (51) and (52):

$$\frac{d\phi_x^l(t)}{dt} = \frac{4T_d\frac{dx_s^l(t)}{dt}}{4T_d^2 + T_e^2 + 4T_e x_s^l(t) + 4x_s^l(t)^2}, \tag{53}$$

$$\frac{d\phi_x^r(t)}{dt} = \frac{4T_d\frac{dx_s^r(t)}{dt}}{4T_d^2 + T_e^2 - 4T_e x_s^r(t) + 4x_s^r(t)^2}, \tag{54}$$

$$\frac{d\phi_y(t)}{dt} = \frac{T_d\frac{dy_s(t)}{dt}}{T_d^2 + y_s(t)^2}. \tag{55}$$

If the camera parameters are known and the position of a moving object in the world space is given by $\mathbf{P}_w(t) = [X_w(t), Y_w(t), Z_w(t)]^\intercal$, (2) and (5) can be used to derive, the angular eye positions over time:

$$\phi_x^l(t) = \arctan\left(\frac{mfT_c + 2mfX_w(t) + T_e Z_w(t)}{2T_d Z_w(t)}\right), \tag{56}$$

$$\phi_x^r(t) = \arctan\left(\frac{-mfT_c + 2mfX_w(t) - T_e Z_w(t)}{2T_d Z_w(t)}\right), \tag{57}$$

$$\phi_y(t) = \arctan\left(\frac{mfY_w(t)}{T_d Z_w(t)}\right). \tag{58}$$

The angular eye velocities can be derived by differentiating (56), (57) and (58) as given by (59)-(61):

$$\frac{d\phi_x^l(t)}{dt} = \frac{2mfT_d\left(2Z_w(t)X_w'(t) - (T_c + 2X_w(t))Z_w'(t)\right)}{m^2f^2T_c^2 + 4m^2f^2X_w(t)^2 + 2mfT_eT_cZ_w(t) + (4T_d^2 + T_e^2)Z_w(t)^2 + 4mfX_w(t)(mfT_c + T_eZ_w(t))}, \tag{59}$$

$$\frac{d\phi_x^r(t)}{dt} = \frac{-2mfT_d\left(2Z_w(t)X_w'(t) + (T_c - 2X_w(t))Z_w'(t)\right)}{m^2f^2T_c^2 + 4m^2f^2X_w(t)^2 + 2mfT_eT_cZ_w(t) + (4T_d^2 + T_e^2)Z_w(t)^2 - 4mfX_w(t)(mfT_c + T_eZ_w(t))}, \tag{60}$$

$$\frac{d\phi_y(t)}{dt} = \frac{mfT_d\left(Z_w(t)Y_w'(t) - Y_w(t)Z_w'(t)\right)}{m^2f^2Y_w(t)^2 + T_d^2Z_w(t)^2}. \tag{61}$$

A few simple cases follow. If the object is moving along the Z_w axis and it is stationary with respect to the other axes, $Z_w(t) = Z_w + v_{wz}t$, $X_w(t) = 0$ $Y_w(t) = 0$ as given by (62)-(64):

$$\frac{d\phi_x^l(t)}{dt} = -\frac{2mfT_dT_cv_{zw}}{m^2f^2T_c^2 + 2mfT_eT_c(Z_w + v_{zw}t) + (4T_d^2 + T_e^2)(Z_w + v_{zw}t)^2}, \tag{62}$$

$$\frac{d\phi_x^r(t)}{dt} = \frac{2mfT_cT_dv_{zw}}{m^2f^2T_c^2 + 2mfT_eT_c(Z_wv_{zw}t) + (4T_d^2 + T_e^2)(Z_w + v_{zw}t)^2}, \tag{63}$$

$$\frac{d\phi_y(t)}{dt} = 0. \tag{64}$$

If the object is moving along the X_w axis and it is stationary with respect to the other axes, $Z_w(t) = Z_w$, $X_w(t) = v_{xw}t$, $Y_w(t) = 0$, the following angular eye velocities result as given by (65)-(67):

$$\frac{d\phi_x^l(t)}{dt} = \frac{4mfT_dv_{xw}Z_w}{m^2f^2T_c^2 + 4m^2f^2v_{xw}^2t^2 + 2mfT_eT_cZ_w + (4T_d^2 + T_e^2)Z_w^2 + 4mfv_{xw}t(mfT_c + T_eZ_w)}, \tag{65}$$

$$\frac{d\phi_x^r(t)}{dt} = \frac{4mfT_dv_{xw}Z_w}{m^2f^2T_c^2 + 4m^2f^2v_{xw}^2t^2 + 2mfT_eT_cZ_w + (4T_d^2 + T_e^2)Z_w^2 - 4mfv_{xw}t(mfT_c + T_eZ_w)}, \tag{66}$$

$$\frac{d\phi_y(t)}{dt} = 0. \tag{67}$$

If the object is moving along the Y_w axis and it is stationary with respect to the other two axes, $Z_w(t) = Z_w$, $X_w(t) = 0$, $Y_w(t) = v_{yw}t$, we have the following angular eye velocities:

$$\frac{d\phi_x^l(t)}{dt} = 0, \tag{68}$$

$$\frac{d\phi_x^r(t)}{dt} = 0, \tag{69}$$

$$\frac{d\phi_y(t)}{dt} = \frac{mfT_d v_{yw} Z_w}{m^2 f^2 v_{yw}^2 t^2 + T_d^2 Z_w^2}. \tag{70}$$

This analysis is important for determining the maximal object speed in the world coordinates or the maximal allowable disparity change, when capturing a fast moving object. If certain angular velocity limits (e.g., 20 deg/sec for ϕ_x [26]) are violated viewer's eyes cannot converge fast enough to follow it, therefore causing visual fatigue. In addition, there are also limits (e.g., 80 deg/sec [27]) for the cases of smooth pursuit (65),(66) and (70) that must not be violated either.

4. Semantic 3D Object Motion Description

In this section, we will present a set of methods for characterizing 3D object motion in stereo video. In our approach, an object (e.g., an actor's face in a movie or the ball in a football game), is represented by a region of interest (ROI), which can be used to refer to an important semantic description regarding object position and motion characterization. It must be noted that, in most cases, neither camera nor viewing parameters are known. In such cases, object motion characterization is based only on object ROI position and motion in the left and right image planes.

Object ROI detection and tracking is overviewed in subsection 4.1. In subsections 4.2 and 4.3, object motion description algorithms are presented, which describe the object motion direction in an object trajectory and the relative motion of two objects, respectively.

4.1. Object Detection and Tracking

We consider that an object is described by a ROI within a video frame or by a ROI sequence, over a number of consecutive frames. These ROIs may be generated by a combination of object detection (or manual initialization) and tracking [28]. Stereo tracking can be performed as well for improved tracking performance [29]. In its

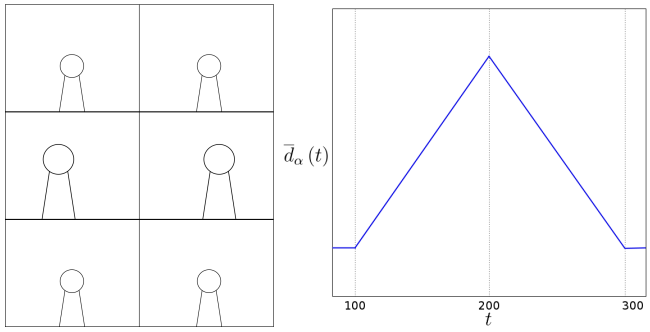
simplest form, a rectangular ROI (bounding box) can be represented by two points $\mathbf{p}_1 = [x_{left}, y_{top}]^\intercal$ and $\mathbf{p}_2 = [x_{right}, y_{bottom}]^\intercal$, where the x_{left}, y_{top}, x_{right} and y_{bottom} are the left, right, top and bottom ROI bounds, respectively. Such ROIs can be found on both the left and right object views. In the case of stereo video, object disparity can be found inside the ROI by disparity estimation [21]. This procedure produces dense or sparse disparity maps [30]. Such maps can be used to obtain an 'average' object disparity, e.g., by averaging the disparity over the object ROI [19]. Alternatively, gross object disparity estimation can be a by-product of the stereo video tracking algorithm, based, e.g., on left/right view SIFT point matching within the left/right object ROIs [31]. In the proposed object motion characterization algorithms, a ROI is represented by its center coordinates $x_{center} = (x_{left} + x_{right})/2$, $y_{center} = (y_{top} + y_{bottom})/2$ along x and y axis, its width and height (if needed) and an overall ('average') disparity value.

In order to better evaluate an overall object disparity value for the object ROI, we first use a pixel trimming process [32], in order to discard pixels that do not belong to the object, since the ROI may contain, apart from the object, background pixels. First, the mean disparity $\overline{d}$ using all pixels inside a central region within the ROI. A pixel within the ROI is retained only when its disparity value is in the range [$\overline{d}$-a,$\overline{d}$+a], where a is an appropriately chosen threshold. Then, the trimmed mean disparity value $\overline{d}_\alpha$ of the retained pixels is computed [19, 32].

4.2. Object motion characterization

In order to characterize object motion, when not knowing the camera and display parameters, we examine the motion separately on x and y axes in the image plane and in the depth space, using object disparities. Specifically, we use the x and y ROI center coordinates $[x_{center}(t), y_{center}(t)]^\intercal$ in both left/right channels and (3) or (7) for characterizing the horizontal and vertical object motion. We can also use the trimmed mean disparity value $\overline{d}_\alpha$ and (3) or (6) for labelling object motion along the depth axis over a number of consecutive video frames. In any case, the unknown parameters are ignored. An example of a $\overline{d}_\alpha$ signal (time series), where t indicates the video frame number is shown in Figure 6. In this particular case, in the theater space the object

first stays at a constant depth Z_d from the viewer, then it moves away and finally it moves closer the viewer. When $\overline{d}_\alpha(t) = 0$, the object is exactly on screen ($Z_d = T_d$). To perform motion characterization, we use first a moving average filter of appropriate length, in order to smooth such a signal over time [33]. Then, the filtered signal can be approximated, using, e.g., a linear piece-wise approximation method [34]. The output of the above process is a sequence of linear segments, where the slope of each linear segment indicates the respective object motion type. The motion duration is defined by the respective linear segment duration. Depending on whether the slope has a negative, positive or close to zero value, respective movement labels can be assigned for each movement, as shown in Table 1. If too short linear segments are found and their slopes are small/moderate, the respective motion characterization can be discarded.



(a) (b)

Figure 6: a) Stereo left/right video frame pairs at times t=100,200,300, b) time series of the trimmed mean object disparity

Table 1: Labels characterizing movement of an object.

Slope value	**negative**	**positive**	**close to zero**
Horizontal movement	left	right	still horizontal
Vertical movement	up	down	still vertical
Movement along the depth axis	backward	forward	still depth

If the stereo camera parameters are known, then the true 3D object position of the left/right ROI center in the world coordinates can be found, using (3) or (16) - (18) for

the object ROI center for the parallel and converging stereo camera setups, respectively. In the uncalibrated case, there are cases where the true 3D object position can be also recovered [35]. The same can be done for the display space, if we know the display parameters m, T_d, T_e, using the ROI center coordinates. Therefore, the movement labels of Table 1 can be used for both world space and display space, following exactly the same procedure for characterizing object motion in the world and display spaces, by using the vector signals $[X_w(t), Y_w(t), Z_w(t)]^\intercal$ and $[X_d(t), Y_d(t), Z_d(t)]^\intercal$, respectively.

In such cases, characterizations of the form 'object moving away/approaching the camera or the viewer' have an exact meaning. Values of $Z_d(t)$ outside the comfort zone [8] indicate stereo visual quality problems. Large slope of $Z_d(t)$ over time, i.e., its derivative exceeding an acceptable threshold $Z'_d(t) > u_d$, can also indicate stereo quality, e.g., eyes convergence problems.

4.3. Motion characterization of object ensembles

Two (or more) objects or persons may approach to (or distance from) each other. For such motion characterizations of object ensembles, we shall examine two different cases, depending on whether camera calibration or display parameters are known or not. If such parameters are not available, 3D world or display coordinates can not computed. Thus, object ensemble motion can be labelled independently along the spatial (image) x, y axes and along the 'depth' axis (using the trimmed average disparity values), only for the parallel camera setup and display. For a number of consecutive video frames, the ROI center coordinates of the left and right video channels are combined into $X^i_{center} = \frac{x^i_{lcenter} + x^i_{rcenter}}{2\left(Te - \overline{d}_{\alpha i}\right)}$ and $Y^i_{center} = \frac{y^i_{center}}{Te - \overline{d}_{\alpha i}}$ (a typical value for T_e is used) using (7) or $X^i_{center} = \frac{x^i_{lcenter} + x^i_{rcenter}}{2\overline{d}_{\alpha i}}$ and $Y^i_{center} = \frac{y^i_{center}}{\overline{d}_{\alpha i}}$ using (3), for the display or parallel camera, respectively, in all cases the unknown parameters are ignored. The Euclidean distances between $\mathbf{p}^i = \left[X^i_{center}, Y^i_{center}\right]^\intercal$ and $\mathbf{p}^j = \left[X^j_{center}, Y^j_{center}\right]^\intercal$ and the respective disparity values $\overline{d}_{\alpha i}$ and $\overline{d}_{\alpha j}$ of two

objects i, j are computed as follows:

$$D_{xy} = \sqrt{(X^i_{center} - X^j_{center})^2 + (Y^i_{center} - Y^j_{center})^2}, \quad (71)$$

$$D_d = \sqrt{(\overline{d}_{\alpha i} - \overline{d}_{\alpha j})^2}. \quad (72)$$

The resulting two signals are filtered and approximated by linear segments, as described in the previous subsection. Similarly, depending on whether the linear segment slope has a negative, positive or close to zero value, the corresponding motion label can be assigned, as shown in Table 2. Even in the absence of camera and display parameters, disparity information can help in inferring the relative motion of two objects: if both D_{xy} and D_d decrease, the objects come closer in the 3D space. However, in such a case no Euclidean distance (e.g., in meters) can be found.

Table 2: Labels characterizing the 3D motion of object ensembles without using calibration/viewing parameters.

Slope value	**negative**	**positive**	**close to zero**
xy movement	approaching_xy	moving_away_xy	equidistant_xy
Depth movement	approaching_depth	moving_away_depth	equidistant_depth

The same procedure can be extended to the case of more than two objects: we can characterize whether their geometrical positions converge or diverge. To do so, we can find the dispersion of their positions vs their center of gravity in the xy domain and in the 'depth' domain:

$$D_{xy} = \sqrt{\sum_{i=1}^{N} \left[(X^i_{center} - \overline{X}_{center})^2 + (Y^i_{center} - \overline{Y}_{center})^2\right]}, \quad (73)$$

$$D_d = \sqrt{\sum_{i=1}^{N} (\overline{d}_{\alpha i} - \overline{\overline{d}}_{\alpha})^2}. \quad (74)$$

and then perform the above mentioned smoothing and linear piece-wise approximation.

When camera calibration parameters are available, the world coordinates $[X_w, Y_w, Z_w]^\intercal$ of an object, which is described by the respective ROI center $[x_{center}, y_{center}]^\intercal$ and trimmed mean disparity value $\overline{d}_\alpha$, can be computed by the equations using (3) and (16),

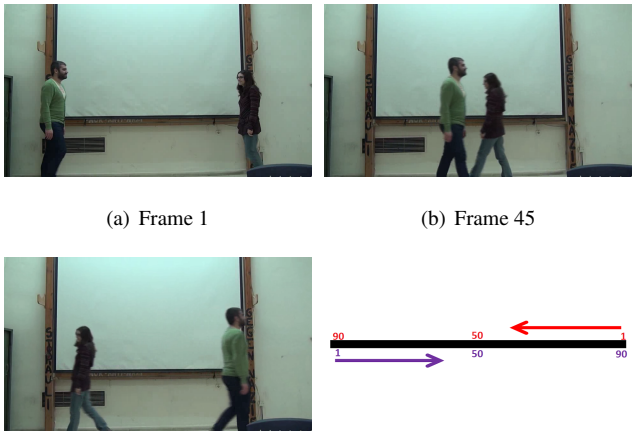
(17), (18) for the parallel and converging camera setup, respectively. Consequently, the actual distance between two objects, which are represented by the two points $\mathbf{P}_1$ and $\mathbf{P}_2$, can be calculated by using the Euclidean distance $\|\mathbf{P}_1 - \mathbf{P}_2\|_2$ in the 3D space. Then, the same approach using smoothing and linear piece-wise approximation can be used for characterizing the motion of two objects.The same procedure can be applied for characterizing their motion in the display space, if the display parameters are known.

5. Experimental Results

5.1. Indoor Scenes

5.1.1. Stereo Dataset Description

For evaluating and assessment the proposed motion labelling methods, we created a set of stereo videos recorded indoors with a stereo camera with known calibration parameters. Specifically, the stereo camera has parallel geometry with a focal length of 34.4 mm and baseline equal to 140 mm. In each video, two persons move along motion trajectories belonging to three different categories. In the first video category the subjects stand facing each other and start walking parallel to the camera, approaching one another up to the middle of the path and then moving away. Figure 7 displays three representative frames of such a stereo video and a diagram (top view), which shows the persons' motion trajectories on the X_wZ_w plane. In the second video category (Figure 8), the persons walk diagonally, following X-shaped paths. Again, the two subjects approach one another during their way up to the middle of the path and then start moving away. In the third video category, the two subjects follow each other on an elliptical path, as depicted in Figure 9. In the beginning, they stand at each end of the major ellipse axis and then start moving clockwise. For a small number of frames their distance is almost constant and their movement can be considered as equidistant. Then, when they come close to the minor ellipse axis, they approach one another and, afterwards, they start moving away again. When reaching again the major ellipse axis, their distance remains almost constant again for a small time period and their movement can



(a) Frame 1 (b) Frame 45

(c) Frame 80 (d) Persons' trajectories. The numbers indicate frames.

Figure 7: Example video frames and respective person's trajectories for the first video category.

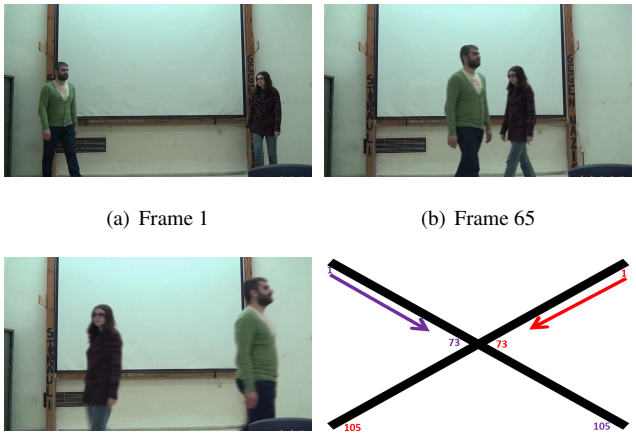
again be considered equidistant. Continuing their movement, they start approaching and then moving away, until they reach their initial positions.

5.1.2. Preprocessing Phase

Before executing the proposed algorithms, a preprocessing step was necessary. First, the disparity maps for each video were extracted. A typical example of a left and right video frame with the respective disparity maps is presented in Figure 10. Next, the ROI trajectories of the two persons were computed. The heads of the two persons were manually initialized at the first frame for each video and were tracked by using the tracking algorithm described in [28]. This process was applied separately on each stereo video channel and the results were copied on the corresponding disparity channels. An example of the tracked person is presented in Figure 11. Finally, for each ROI, the corresponding ROI center coordinates and trimmed average disparity value $\overline{d}_{\alpha}$ were computed, as described in subsection 4.1.

5.1.3. Movement Description Examples

For the three videos depicted in Figures 7-9, the algorithm for movement characterization described in 4.2 was performed. In Table 3, the generated video segments with



(a) Frame 1 (b) Frame 65

(c) Frame 95 (d) Persons' trajectories. The numbers indicate frames.

Figure 8: Example video frames and respective person's trajectories for the second video category.

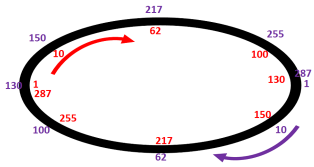
the corresponding horizontal motion label of the man and woman are shown. The ROI center x coordinates of the man and woman and the output of the linear approximation process for the video depicted in Figure 9 are shown in Figures 12 and 13 respectively. If no disparity is used, it seems that the persons meet twice approximately at video frames 60 and 210. This is not the case, since their disparities differ at the respective times, as shown in Figure 13.

The output of the proposed algorithm for characterizing the relative motion between two objects, with known calibration parameters, for the three videos shown in Figures 7 , 8 and 11, are depicted in Figures 14, 15 and 16, respectively. Distance are now measured in meters in the world space. As shown in Figure 14, two subjects are approaching in the video frame interval [1,48], are equidistant in the interval [49,56] and are moving away in the interval [57,90]. Similarly, the result of algorithm for the video depicted in Figure 8 and shown in Figure 15 is that two subjects approach in the frame interval [1,71], are equidistant in the interval [72,75] and move away in the interval [76,105]. The generated labels for the last video are shown in Table 4, the two subjects are equidistant in the interval [1,7], are approaching in the interval [8,61] and are moving away in the interval [62,93]. The same motion pattern is repeated in the

(a) Frame 10

(b) Frame 65

(c) Frame 185



(d) Persons' trajectories. The numbers indicate frames.

Figure 9: Example video frames and respective person's trajectories for the third video category.

(a) Left frame

(b) Right frame

(c) Left disparity map

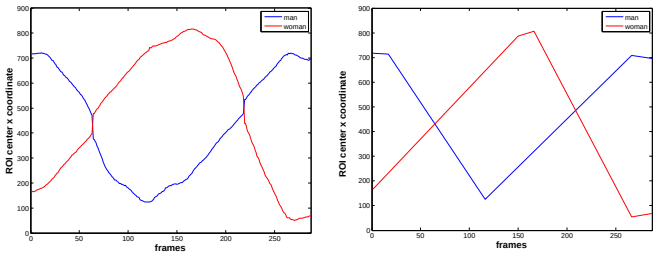
(d) Right disparity map

Figure 10: Sample video frames with their disparity maps.

frame intervals [94,152], [153,216], [217,261]. Finally, the two subjects are equidistant again in [262,285].

(a) Left frame (b) Right frame

Figure 11: Sample video frames with ROIs.



(a) (b)

Figure 12: a) x coordinate of the ROI center of woman and man for the video depicted in Figure 9 and (b) the result of linear approximation.

5.2. Outdoor/challenging scenes and quantitative performance evaluation

In order to assess the robustness of the presented motion labelling methods in real conditions, we created a set of videos recorded outdoors with the same stereo camera in realistic conditions. These videos depict walking humans and moving cars. As shown in Figure 17, where some representative frames are displayed, the background is quite complex and lighting conditions are far from being ideal. The type of motion of the tracked object(s) was manually labelled on these videos so as to create ground-truth labels. The number of the instances for each different motion type appearing in these videos are given in Table 5. As in previous section, the disparity maps were extracted, while the ROI trajectories of the various subjects, namely humans and cars, were computed by a combination of manual initialization and automatic tracking.

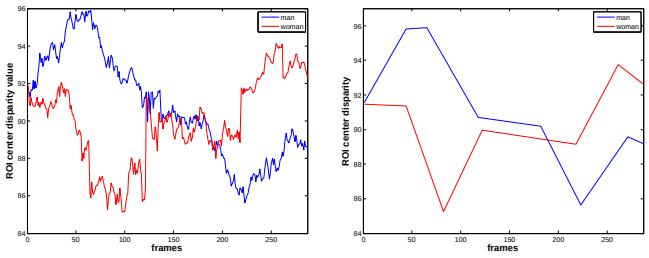
The algorithms for movement characterization and for characterizing the relative motion between two objects on videos captured with known calibration parameters (Subsection 5.1.1) were applied on these videos. Table 6 shows the mean temporal

Table 3: The generated man/woman labels.

Video type	Person	Start frame	End frame	Label
a	man	1	90	right
b	man	1	105	right
c	man	1	17	still horizontal
c	man	18	116	left
c	man	117	266	right
c	man	267	287	still horizontal
a	woman	1	90	left
b	woman	1	105	left
c	woman	1	150	right
c	woman	151	166	still horizontal
c	woman	167	265	left
c	woman	266	287	still horizontal

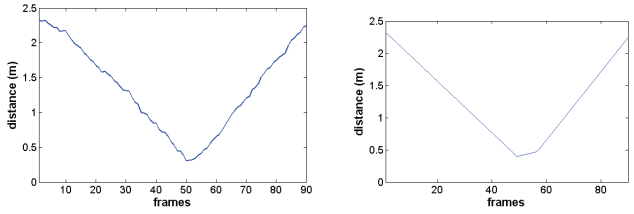
The generated labels for motion characterization
for the videos shown in Figure 7 (a), Figure 8 (b) and Figure 9 (c).

overlap between the predicted labels (each corresponding to a motion segment i.e. a number of frames) and ground-truth labelled motion segments for each different motion type. As can be seen, a high accuracy is achieved for most motion types, proving the effectiveness and robustness of the proposed method in real world stereo videos. For example, an accuracy bigger that 91% was achieved in the case of motion types/labels "left", "right", "still horizontal", "still depth", "still vertical", "approaching", "moving away" and "equidistant". On the other hand, smaller but still fairly good accuracies can be noted for other motion types/labels related to motion along depth and the vertical direction, namely "forward", "backward", "up", "down". For the "forward"/"backward" motion, this can be explained by the fact that disparity is not very accurate especially in image parts with big depth. For the motion along the vertical axis ("up"/"down") errors can be explained by the fact that in these instances the subject is mainly moving along the depth axis, and only slightly in the vertical axis. Thus, the corresponding po-



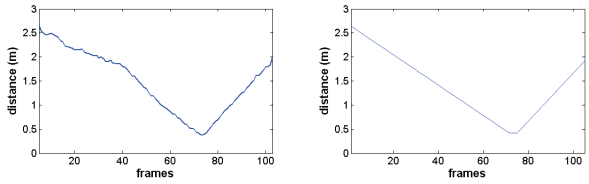
(a) (b)

Figure 13: a) Trimmed average disparity of the woman and man ROI for the video depicted in Figure 9, b) the result of linear approximation.



(a) (b)

Figure 14: a) Person distances (in meters) calculated in the 3D space, b) the result of linear approximation of the distance signal for the video depicted in Figure 7.

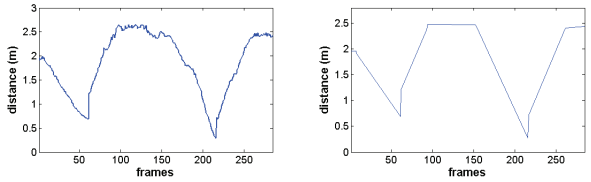


(a) (b)

Figure 15: a) Person distances (in meters) calculated in the 3D space, b) the result of linear approximation of the distance signal for the video depicted in Figure 8.

sition signal has a small slope resulting in some cases false predicted labels, i.e. “still vertical” instead of “up”/“down”.

Finally, Figure 18 exemplifies the importance of applying an appropriate filter to the signal representing the position of an object or the distance between two objects



(a) (b)

Figure 16: a) Person distances (in meters) calculated in the 3D space, b) the result of linear approximation of the distance signal for the video depicted in Figure 9.

Table 4: The generated motion labels for the video depicted in Figure 9.

Start frame	End frame	Label
1	7	equidistant
8	61	approaching
62	93	moving away
94	152	equidistant
153	216	approaching
217	261	moving away
262	285	equidistant

over time, towards overcoming possible tracking failures caused e.g., by occlusion. Figure 18(b) shows the predicted labels of the position along depth of a face tracked over time with and without filtering, where for some frames (Figure 18(a)) the face has been mis-tracked due to occlusion by another face. As can be seen, the predicted labels when applying filtering are in agreement with the ground-truth ones, In contrast, when no filtering is applied, two small segments are given false labels.

6. Conclusion

In this paper, 3D object motion mapping from the world space to the image space and to the display (theater) space is first analysed in a novel way. The effect of screen disparity changes on the viewing experience is presented. Then new algorithms are presented that characterize object motion in stereo video content along the horizontal,

(a) (b)

(c) (d)

Figure 17: Example video frames.

Table 5: Number of instances for each different motion type (label).

Motion label	#	Motion label	#
left	25	up	4
right	14	down	5
still horizontal	6	still vertical	24
forward	5	approaching	12
backward	6	moving away	15
still depth	29	equidistant	6

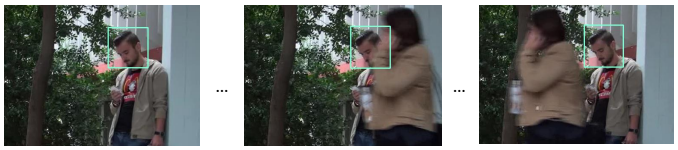
vertical and depth axis and assign labels depending on whether two objects approach each other or move away. On the other hand, a mathematical analysis is presented about the relation of object motion in world coordinates compared to their perceived motion in the display (theater) space. Finally, we examine whether and how the viewing experience is affected by disparity manipulations.

7. Acknowledgement

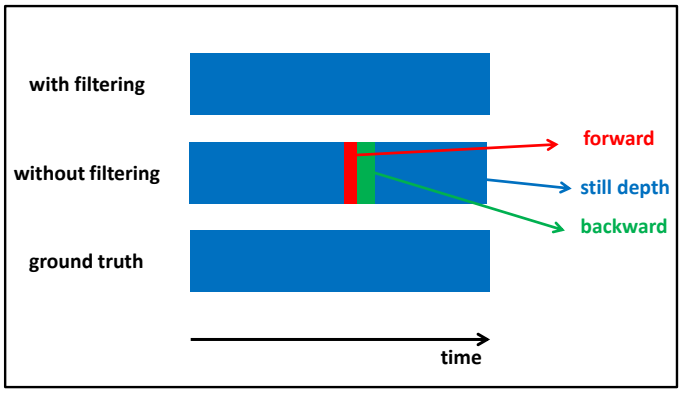
The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287674

Table 6: Mean overlap (%) for each different motion type.

Motion label	#	Motion label	#
left	94.73	up	75.98
right	94.74	down	78.63
still horizontal	100.00	still vertical	99.20
forward	70.58	approaching	97.50
backward	73.92	moving away	93.32
still depth	99.93	equidistant	90.91



(a) Face ROIs on sample frames



(b) predicted labels

Figure 18: The effect of filtering on a trajectory where occlusion occurs.

(3DTVS). This publication reflects only the author's views. The European Union is not liable for any use that may be made of the information contained therein.

Appendix A. Calculation of World Coordinates in Converging Camera Setup Geometry

The auxiliary angles, which are shown in Figure A.19, can be expressed as:

$$\psi_l = \arctan\left(\frac{x_c^l}{f}\right), \tag{A.1}$$

$$\psi_r = \arctan\left(\frac{x_c^r}{f}\right), \tag{A.2}$$

$$\phi_l = \frac{\pi}{2} - \psi_l - \theta, \tag{A.3}$$

$$\phi_r = \frac{\pi}{2} - \psi_r - \theta, \tag{A.4}$$

$$\omega = \pi - \phi_l - \phi_r = \psi_l + \psi_r + 2\theta. \tag{A.5}$$

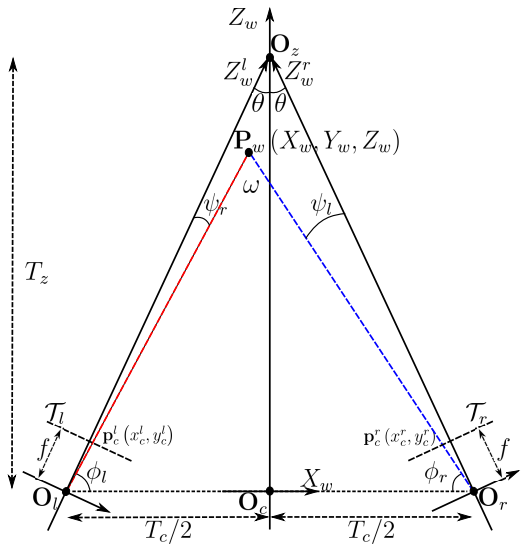


Figure A.19: Converging stereo camera setup geometry.

The law of sines in the triangle $(\widehat{O_l P_w O_r})$ gives us:

$$\frac{T_c}{\sin\omega} = \frac{(\overline{P_w O_l})}{\sin\phi_r} = \frac{(\overline{P_w O_r})}{\sin\phi_l}. \tag{A.6}$$

Thus, Z_w can be expressed as:

$$Z_w = \overline{(P_w O_l)}\sin\phi_l = \overline{(P_w O_r)}\sin\phi_r = T_c\frac{\sin(\phi_l)\sin(\phi_r)}{\sin\omega}. \tag{A.7}$$

After replacing ω, ϕ_l and ϕ_r, (A.7) is simplified as follows:

$$\begin{aligned} Z_w &= T_c \frac{\sin\left(\frac{\pi}{2} - \psi_l - \theta\right)\sin\left(\frac{\pi}{2} - \psi_r - \theta\right)}{\sin\left(\psi_l + \psi_r + 2\theta\right)} = \\ &= T_c \frac{f - \left(x_c^l - x_c^r + \frac{x_c^l x_c^r}{f}\tan\theta\right)\tan\theta}{x_c^l - x_c^r + \left(2f + 2\frac{x_c^l x_c^r}{f} - x_c^l\tan\theta + x_c^r\tan\theta\right)\tan\theta} \end{aligned} \tag{A.8}$$

The equations (16) and (17) can be proved with the same methodology:

$$X_w = \overline{P_w O_l} cos\phi_l - \frac{T_c}{2}. \tag{A.9}$$

Y_w can be obtained by projecting P_w on the left optical axis and then using triangle similarities:

$$Y_w = \frac{\overline{P_w O_l} y_c^l}{f} cos\psi_l. \tag{A.10}$$

References

[1] A. Smolic, P. Kauff, S. Knorr, A. Hornung, M. Kunter, M. Muller, M. Lang, Three-Dimensional Video Postproduction and Processing, Proceedings of the IEEE 99 (4) (2011) 607–625.

[2] F. I. Bernard F. Coll, K. O'Connell, 3dtv at home: status, challenges and solutions for delivering a high quality experience, in: Proceedings of the Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2010.

[3] A. Hilton, J. Y. Guillemaut, J. Kilner, O. Grau, G. Thomas, 3D-TV Production From Conventional Cameras for Sports Broadcast, IEEE Transactions onBroadcasting 57 (2) (2011) 462–476.

[4] J. DeFilippis, 3D Sports Production at the London 2012 Olympics, SMPTE Motion Imaging Journal 122 (1) (2013) 20–23.

[5] R. Mintz, S. Litvak, Y. Yair, 3D-Virtual Reality in Science Education: An Implication for Astronomy Teaching, Journal of Computers in Mathematics and Science Teaching 20 (3) (2001) 293–305.

[6] M. Zyda, From visual simulation to virtual reality to games, Computer 38 (9) (2005) 25–32.

[7] A. Smolic, 3D video and free viewpoint video-From capture to display, Pattern Recognition 44 (9) (2011) 1958–1968.

[8] B. Mendiburu, 3D Movie Making - Stereoscopic Digital Cinema from Script to Screen., Focal Press, 2009.

[9] S. Koppal, C. Zitnick, M. Cohen, S. B. Kang, B. Ressler, A. Colburn, A Viewer-Centric Editor for 3D Movies, Computer Graphics and Applications, IEEE 31 (1) (2011) 20–35.

[10] E. Larsen, P. Mordohai, M. Pollefeys, H. Fuchs, Temporally consistent reconstruction from multiple video streams using enhanced belief propagation, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, 2007, pp. 1–8.

[11] M. Yang, X. Cao, Q. Dai, Multiview video depth estimation using spacial-temporal consistency, in: Proceedings of the British Machine Vision Conference, BMVA Press, 2010, pp. 67.1–67.11.

[12] H. Jiang, H. Liu, P. Tan, G. Zhang, H. Bao, 3d reconstruction of dynamic scenes with multiple handheld cameras, in: A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, C. Schmid (Eds.), Computer Vision ECCV 2012, Springer Berlin Heidelberg, 2012, pp. 601–615.

[13] S. Jeannin, A. Divakaran, Mpeg-7 visual motion descriptors, IEEE Transactions on Circuits and Systems for Video Technology 11 (6) (2001) 720–724.

[14] E. Stone, M. Skubic, Evaluation of an inexpensive depth camera for passive in-home fall risk assessment, in: 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011, pp. 71–77.

[15] K. E. Ozden, K. Cornelis, L. V. Eycken, L. V. Gool, Reconstructing 3d trajectories of independently moving objects using generic constraints, Computer Vision and

Image Understanding 96 (3) (2004) 453 – 471, special issue on model-based and image-based 3D scene representation for interactive visualization.

[16] C. Yuan, G. Medioni, 3d reconstruction of background and objects moving on ground plane viewed from a moving camera, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, 2006, pp. 2261–2268.

[17] D. Zou, Q. Zhao, H. S. Wu, Y. Q. Chen, Reconstructing 3d motion trajectories of particle swarms by global correspondence selection, in: IEEE 12th International Conference on Computer Vision, 2009, pp. 1578–1585.

[18] F. Speranza, W. J. Tam, R. Reunaud, N. Hur, Effect of Disparity and Motion on Visual Comport of Stereoscopic Images, in: Proc. SPIE 6055, Stereoscopic Displays and Virtual Reality Systems XIII, 2006.

[19] N. Papanikoloudis, S. Delis, N. Nikolaidis, I. Pitas, Semantic description in stereo video content for surveillance applications, in: Biometrics and Forensics (IWBF), 2013 International Workshop on, 2013, pp. 1–4.

[20] T. Theodoridis, K. Papachristou, N. Nikolaidis, I. Pitas, Object motion description in stereoscopic videos, in: 3D Imaging (IC3D), 2013 International Conference on, 2013, pp. 1–7.

[21] E. Trucco, A. Verri, Introductory Techniques for 3-D Computer Vision, Prentice Hall PTR, 1998.

[22] A. Woods, T. Docherty, R. Koch, Image distortions in stereoscopic video systems, in: Proceedings of the SPIE Volume 1915, Stereoscopic Displays and Applications IV, 1993.

[23] C. MacLachlan, H. C. Howland, Normal values and standard deviations for pupil diameter and interpupillary distance in subjects aged 1 month to 19 years, Ophthalmic and Physiological Optics 22 (3) (2002) 175–182.

[24] W. Robinett, J. P. Rolland, Computational model for the stereoscopic optics of a head-mounted display, in: Stereoscopic Displays and Applications II, Proc. SPIE, Vol. 1457, 1991, pp. 140–160.

[25] L.-F. Cheong, X. Xiang, What do we perceive from motion pictures? a computational account, J. Opt. Soc. Am. A 24 (6) (2007) 1485–1500.

[26] I. P. Howard, B. J. Rogers, Binocular Vision and Stereopsis, Oxford University Press, 1996.

[27] C. H. Meyer, A. G. Lasker, D. A. Robinson, The upper limit of human smooth pursuit velocity , Vision Research 25 (4) (1985) 561 – 563.

[28] O. Zoidi, A. Tefas, I. Pitas, Visual Object Tracking Based on Local Steering Kernels and Color Histograms, IEEE Transactions on Circuits and Systems for Video Technology 23 (5) (2013) 870–882.

[29] O. Zoidi, N. Nikolaidis, I. Pitas, Appearance based object tracking in stereo sequences, in: 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[30] D. Scharstein, R. Szeliski, A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms, International Journal on Computer Vision 47 (1-3) (2002) 7–42.

[31] G. Chantas, N. Nikolaidis, I. Pitas, A bayesian methodology for visual object tracking on stereo sequences, in: 11th IEEE IVMSP Workshop: 3D Image/Video Technologies and Applications, 2013, pp. 1–4.

[32] I. Pitas, A. Venetsanopulos, Nonlinear Digital Filters, Boston: Kluwer, 1990.

[33] A. V. Oppenheim, R. W. Schafer, Digital Signal Processing, Prentice Hall, 1975.

[34] I. Pitas, Digital Image Processing Algorithms, Prentice Hall, 1993.

[35] G. Xu, Z. Zhang, Epipolar Geometry in Stereo, Motion, and Object Recognition: A Unified Approach, Kluwer Academic Publishers, Norwell, MA, USA, 1996.