OBJECT MOTION DESCRIPTION IN STEREOSCOPIC VIDEOS

T. Theodoridis, K. Papachristou, N. Nikolaidis and I. Pitas

Aristotle University of Thessaloniki
Department of Informatics
Box 451
54124 Thessaloniki, Greece
email: {nikolaid,pitas}@aiia.csd.auth.gr

ABSTRACT

The efficient search and retrieval of the increasing volume of stereoscopic videos drives the need for the semantic description of its content. The derivation of disparity (depth) information from stereoscopic content allows the extraction of semantic information that is inherent to 3D. The purpose of this paper is to propose algorithms for semantically characterizing the motion of an object or groups of objects along any of the X, Y, Z axes. Experimental results are also provided.

Index Terms— Semantic labelling, stereo video, motion characterization.

1. INTRODUCTION

In the recent years, the number of produced 3D movies and 3D (stereoscopic) video content in general has been growing significantly. Indeed, a large number of 3D movies have been released and some of them, such as Avatar [11] were huge box-office hits. This, along with the increasing penetration of 3DTV sets in the market, have boosted the delivery of 3D productions, such as movies and documentaries to home through 3DTV channels or Blu-ray disks [3]. Since 3DTV content is now widely available, it must be semantically described towards its archiving, fast search and retrieval. Object motion characterization, in the world (video acquisition) space is an important type of semantic description that one can derive. In this paper, we focus on 3D motion description in stereo video content and propose algorithms for semantic labelling of human, object or object groups motion. Typically, stereo video is shot with a stereo camera, whose parameters include the focal length and the baseline [5], and displays objects residing and moving in the world space (X_w, Y_w, Z_w). We utilize the depth information which is implicitly available through disparity estimation between the left and right views and examine various cases, where camera calibration information may or may not be available. For example, we can characterize video segments where an object approaches the camera or where two objects approach each other in the real world. Such characterization is not possible in classical single view video, without resorting to other side information to get 3D position/motion clues [5]. The derived semantic description is useful in various applications, such as 3D/3DTV video archival and retrieval. The paper extends the work in [7] by including the study of motion captured by calibrated cameras.

The rest of paper is organized as follows. In section 2, algorithms for characterizing object and object groups motion are proposed. In section 3, experimental results for motion characterization are presented. Finally, in section 4, concluding remarks are given.

2. SEMANTIC OBJECT MOTION DESCRIPTION

In this section, we will present a set of methods for characterizing 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 two regions of interest (ROI, bounding box) in the left and right video frames. These ROIs may be generated in every frame by a combination of object detection (or manual initialization) and tracking [14]. Stereo tracking can be performed as well for improved performance [13]. A rectangular ROI can be represented by two points $\mathbf{p}_1\,[x_{left}, y_{top}]^\intercal$ and $\mathbf{p}_2\,[x_{right}, y_{bottom}]^\intercal$ namely its upper left and lower right corners. It must be noted that, in most cases, camera parameters are unknown. In such cases, object motion characterization is based only on object ROI position and motion in the left and right image planes.

Object disparity can be evaluated inside the ROI by using a disparity estimation algorithm [12] that generates dense or sparse disparity maps [10]. Such maps can be used to obtain an 'average' object disparity, e.g. by averaging image disparity over the object ROI [7]. Alternatively, gross object disparity estimation can be a byproduct of the tracking algorithm based e.g. on left/right view SIFT point correspondences [2].

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

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

In the proposed object motion characterization algorithms, a ROI is represented by its center coordinates along x and y axis, its width and height (if needed) and an overall ('average') disparity value. In order to obtain a less noisy overall object disparity value from the object ROI, we first use a pixel trimming process [9], in order to discard pixels that do not belong to the object, since the ROI may contain, apart from the object, background pixels. Pixel trimming first computes the mean disparity d_m using all pixels inside a central region within the ROI. A pixel within the ROI is retained for subsequent calculations, only when its disparity value is in the range [d_m-a,d_m+a], where a is an appropriately chosen threshold. Our experiments showed that a value of $a = 5$ is effective for most small sized objects such as faces. Then, the trimmed mean disparity value $\overline{d}_\alpha$ of the retained pixels is computed [7, 9].

2.1. Object motion characterization

We examine the case of a parallel camera setup. If the camera parameters (the focal length f and baseline T_c) are known, the world space coordinates (X_w, Y_w, Z_w) of a point can be recovered from its left ($\mathbf{p}_c^l = \left[x_c^l, y_c^l\right]^\intercal$) and right ($\mathbf{p}_c^r = [x_c^r, y_c^r]^\intercal$) frame projections, as follows [12]:

$$Z_w = -\frac{fT_c}{d_c}, \tag{1}$$

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

where $d_c = x_c^r - x_c^l$ is the stereo disparity. If the stereo camera parameters are known, then the true 3D object position in the world coordinates can be found, using (1), (2) for the object ROI center. In order to characterize object motion with unknown camera parameters, they are ignored, i.e. both f and T_c are set to 1 since they do not affect the direction of motion in each axis. Furthermore, we examine the motion separately on the x and y axes in the image plane and in the depth space using object disparities. Specifically, we use the x and y coordinates of the ROI center $[x_{center}(t), y_{center}(t)]^\intercal$ in both channels within (2) for evaluating horizontal and vertical object position and the trimmed mean disparity value $\overline{d}_\alpha$ within (1) for evaluation the object's position along the depth axis over a number of consecutive video frames. To perform motion characterization, we use first a moving average filter of appropriate length, in order to smooth the signals $X_w(t)$, $Y_w(t)$, $Z_w(t)$ over time [6]. Then, the filtered signal is approximated, using, e.g., a linear piece-wise approximation method [8]. 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, such as left/right, up/down or backward/forward (in depth) movement. 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. The duration of a specific motion type is defined by the respective linear segment duration. If too short linear segments are found (a few frames) they can be regarded as noise.

2.2. Motion characterization of object groups

Two (or more) objects or persons may approach or move away from each other. For the motion characterization of groups of objects, we shall examine two different cases, depending on whether camera calibration data are known or not.

2.2.1. Uncalibrated cameras

If camera parameters are not available, 3D world coordinates can not be computed. Thus, group object motion can only be labelled independently along the spatial (image) x, y axes and along the 'depth' axis (using the trimmed average disparity values). For the ith object and a number of consecutive video frames $(1 \dots N)$, the ROI center coordinates of the left and right channels are combined into:

$$X_{center}^i(t) = \frac{x_{lcenter}^i(t) + x_{rcenter}^i(t)}{2\overline{d}_{\alpha i}},$$
$$Y_{center}^i(t) = \frac{y_{center}^i(t)}{\overline{d}_{\alpha i}}.$$

The Euclidean distances between two objects i, j located at $p^i = \left[X_{center}^i, Y_{center}^i\right]^\intercal$ and $p^j = \left[X_{center}^j, Y_{center}^j\right]^\intercal$ and having the respective disparity values $\overline{d}_{\alpha i}$ and $\overline{d}_{\alpha j}$ are computed as follows:

$$D_{xy}(t) = \Big[(X_{center}^i(t) - X_{center}^j(t))^2 + (Y_{center}^i(t) - Y_{center}^j(t))^2\Big]^{1/2}, \tag{3}$$

$$D_d(t) = \left[(\overline{d}_{\alpha i}(t) - \overline{d}_{\alpha j}(t))^2\right]^{1/2}. \tag{4}$$

The resulting two signals are filtered and approximated by linear segments, as described in the previous subsection.

Table 2. Labels characterizing the 3D motion of object ensembles without using calibration 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

Similarly, depending on whether the linear segment slope has a negative, positive or close to zero value, a movement label, such as approaching or moving away in the xy (image) plane, can be assigned, as shown in Table 2.

Even in the absence of camera parameters, disparity information can help in inferring the relative motion of two objects in the 3D space in certain cases: if both D_{xy} and D_d decrease/increase, the objects come closer/move away in the 3D space. However, in such a case no Euclidean distance (e.g. in meters) can be found.

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 overall center of gravity in the xy domain ($\overline{X}_{center}$ and $\overline{Y}_{center}$) and in the 'depth' domain ($\overline{\overline{d}}_{\alpha}$):

$$D_{xy}(t) = \left[\sum_{i=1}^{N}\left[\left(X^i_{center}(t) - \overline{X}_{center}(t)\right)^2 + \left(Y^i_{center}(t) - \overline{Y}_{center}(t)\right)^2\right]\right]^{1/2}, \quad (5)$$

$$D_d(t) = \left[\sum_{i=1}^{N}(\overline{d}_{\alpha i}(t) - \overline{\overline{d}}_{\alpha}(t))^2\right]^{1/2}. \quad (6)$$

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

2.2.2. Calibrated cameras

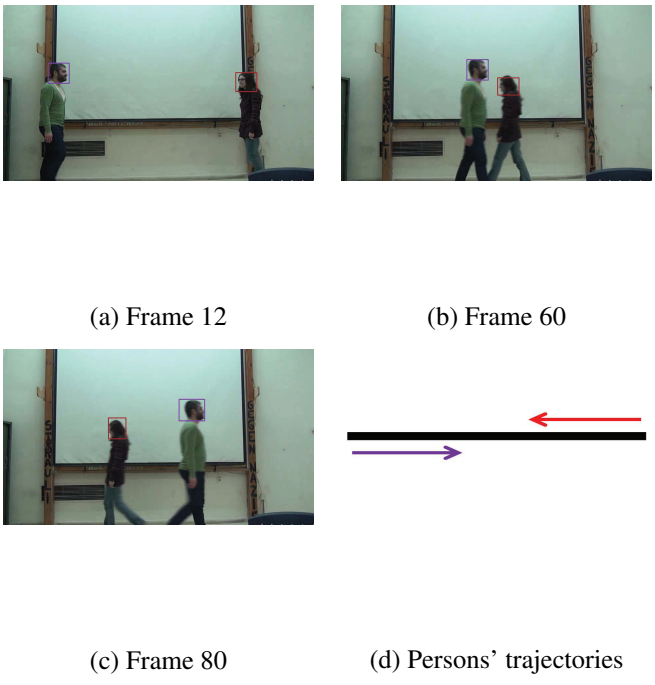
When camera calibration parameters are available, the world coordinates $[X_w, Y_w, Z_w]^\intercal$ of an object, that is described by the respective ROI center $[x_{center}, y_{center}]^\intercal$ and trimmed mean disparity value $\overline{d}_{\alpha}$, can be computed by using (1) and (2). Consequently, the actual distance between two objects, which are represented by the two points $\left[X^1_w, Y^1_w, Z^1_w\right]$ and $\left[X^2_w, Y^2_w, Z^2_w\right]$, can be calculated by using the Euclidean distance in the 3D space:

$$D(t) = \left[(X^1_w(t) - X^2_w(t))^2 + (Y^1_w(t) - Y^2_w(t))^2 + (Z^1_w(t) - Z^2_w(t))^2\right]^{1/2}. \quad (7)$$

Then, the same approach using smoothing and linear piecewise approximation can be used for characterizing the motion of two objects with labels "approaching", "moving away" and "equidistant" when the slope of D is negative, positive and zero, respectively.

3. EXPERIMENTAL RESULTS

To evaluate the proposed method we performed experiments on a set of stereo videos recorded with a stereo camera with known calibration parameters, i.e. a camera with parallel geometry, a focal length of 34.4 mm and baseline equal to 140 mm. In each video two persons are walking. The produced dataset consists of three different categories of persons' trajectories.



(a) Frame 12 (b) Frame 60

(c) Frame 80 (d) Persons' trajectories

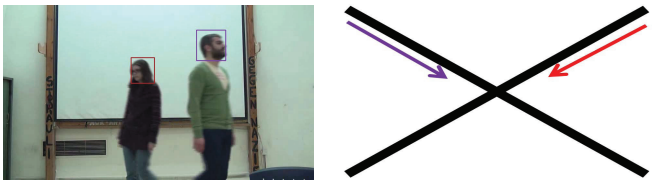
Fig. 1. Example frames and the trajectories of the persons for the first category of videos.

In the first category of videos 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. In Figure 1, three representative frames of such a video and a diagram which shows the trajectories of the persons on the xz plane (top view) are displayed.

In the second category of videos (Figure 2), the persons walk diagonally, following an X-shaped path. Again, the two subjects are approaching one another on their way up to the

middle of the path and then start moving away.

(a) Frame 12 (b) Frame 60



(c) Frame 80 (d) Persons' trajectories

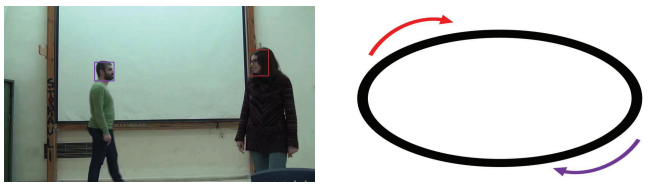
Fig. 2. Example frames and the trajectories of the persons for the second category of videos.

In the third category of videos, the subjects follow an elliptical path as depicted in Figure 3. In the beginning they stand at the major axis of the ellipse and start moving clockwise. For a small number of frames their distance is almost constant and their movement can be considered as equidistant. Then, moving up to the minor axis of the ellipse they are approaching one another and afterwards they start moving away. Reaching again the major axis their distance remains almost constant again for a small time period and their movement can again be considered equidistant. Continuing their movement they start approaching and then moving away until they reach their initial positions.

Disparity maps for each of the videos described above were extracted by using the algorithm [4] that is part of the OpenCV library [1]. A typical example of a left and right frame of the video with the respective disparity maps is presented in Figure 4. The two persons were tracked by using the tracking algorithm described in [13], applied separately on each channel.

The unknown camera parameters case was examined by performing single object motion characterization on the woman in the video depicted in Figure 3. The focal length and camera baseline values were set to 1 ($f = 1$ and $T_c = 1$). The results for the characterization along the X axis are shown in Figure 5.

(a) Frame 20 (b) Frame 70



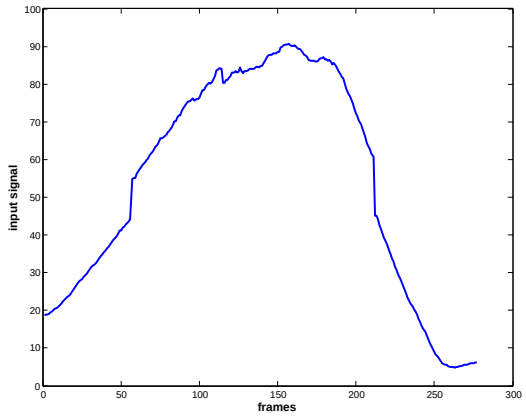
(c) Frame 180 (d) Persons' trajectories

Fig. 3. Example frames and the trajectories of the persons for the third category of videos.

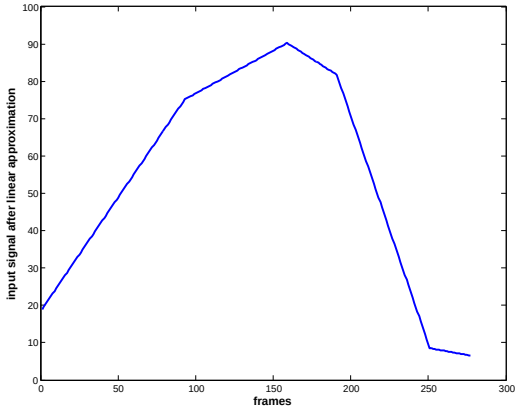
(a) Left frame (b) Right frame

(c) Left disparity maps (d) Right disparity maps

Fig. 4. Sample video frames and their disparity maps.



(e)

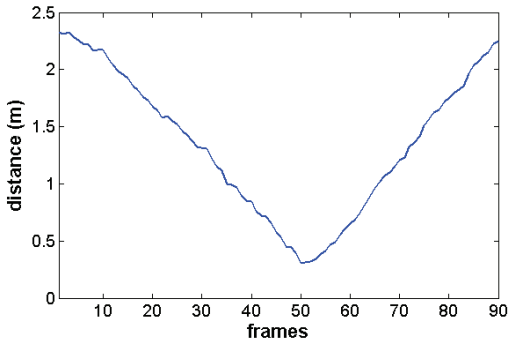


(f)

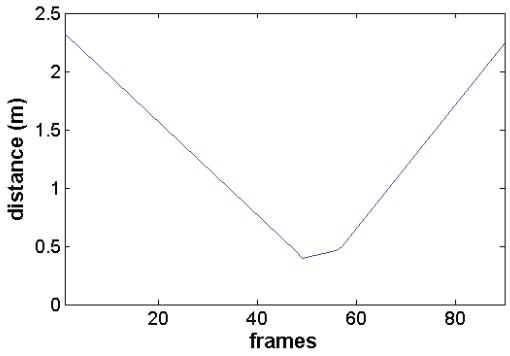
Start frame	End frame	Label
1	93	right
94	189	still
190	252	left
253	285	still

(g)

Fig. 5. (a) Trajectory calculated by our method , (b) the result of linear approximation, and (c) the generated labels for the movement along the X axis for the video depicted in Figure 3.



(a)



(b)

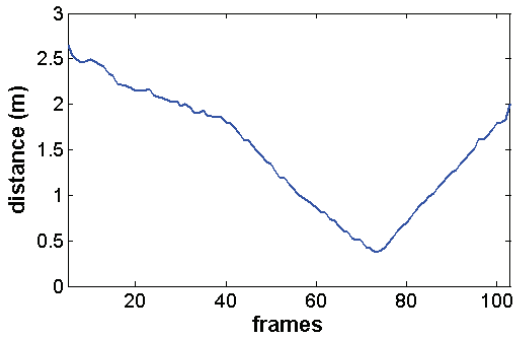
Start frame	End frame	Label
1	48	approaching
49	56	equidistant
57	90	moving away

(c)

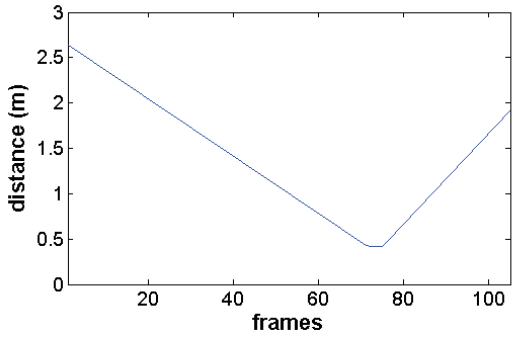
Fig. 6. (a) Distances in 3D space calculated by our method, (b) the result of linear approximation, and (c) the generated labels for the video depicted in Figure 1.

For all three videos depicted in Figures 1 - 3 object motion characterization with known parameters was performed.

For the video in Figure 1 the distances calculated, the output of the linear approximation process and the derived labels are shown in Figure 6. Every line segment represents one of the three possible relative movements, “approaching”, “moving away” and “equidistant”. Results for videos in Figures 2 and 3 are shown in Figures 7 and 8 respectively.



(a)



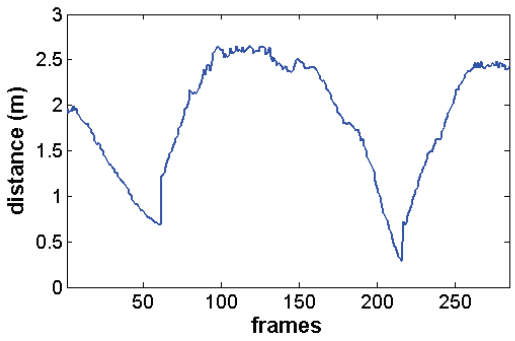
(b)

Start frame	End frame	Label
1	71	approaching
72	75	equidistant
76	105	moving away

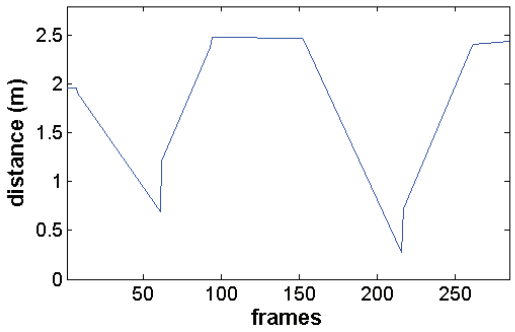
(c)

Fig. 7. (a) Distances in 3D space calculated by our method, (b) the result of linear approximation, and (c) the generated labels for the video depicted in Figure 2.

By comparing the movement of persons in the three different cases with the computed results one can see that the method works properly.



(a)



(b)

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

(c)

Fig. 8. (a) Distances in 3D space calculated by our method, (b) the result of linear approximation, and (c) the generated labels for the video depicted in Figure 3.

4. CONCLUSION

In this paper, an algorithm is presented that characterizes an object's motion in stereo video content along the horizontal, vertical and depth axis and assigns labels such as moving left/right or backwards/forward. In addition, approaches for characterizing the relative movement (approaching, moving away) of objects in the 3D space (when calibration parameters are available) or in the xy (image) plane and the depth dimension (when no calibration parameters are available) are presented. The proposed algorithms make use of disparity information derived from the stereoscopic videos.

5. REFERENCES

[1] G. Bradski, A. Kaehler, and V. Pisarevsky. Learning-based computer vision with intels open source computer vision library. *Intel Technology Journal*, 9(2):119–130, 2005.

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

[3] B. F. Coll, F. Ishtiaq, and 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.

[4] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *Proceedings of IEEE Conference in Computer Vision, vol. 1*, pages 508–515, 2001.

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

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

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

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

[9] I. Pitas and A.N. Venetsanopulos. *Nonlinear Digital Filters*. Boston: Kluwer, 1990.

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

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

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

[13] O. Zoidi, A. Tefas, and I. Pitas. Appearance based object tracking in stereo sequences. In *38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

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