MULTIVIEW VIDEO DATASETS FOR EVALUATION AND TESTING OF 3D TRACKING AND 3D RECONSTRUCTION ALGORITHMS

G. Stamou[†]*, Z. Cernekova*[†]*, N. Nikolaidis*[†] *and A. Sadka*[‡]

[†] Department of Informatics
Aristotle University of Thessaloniki, Box 451, 54124 Thessaloniki, GREECE
Email: {*gstamou,zuzana,nikolaid*}*@aiia.csd.auth.gr*

[‡] Centre for Communication Systems Research
Guildford, Surrey GU2 7XH, United Kingdom
Email: A.Sadka@surrey.ac.uk

ABSTRACT

This paper presents a dataset that can be used for testing and evaluation of computer vision-based algorithms for 3D reconstruction and 3D person tracking. Additional possible uses include the building of 3D human head models and the production of stereoscopic sequences. A number of different scenes are included in the dataset. They are mostly single-subject scenes captured with two different lighting conditions (optimal and sub-optimal), subject motion based on simple as well as random motion trajectories, different distances from the camera and occlusion. The dataset incorporates video data captured simultaneously from a system of 2 cameras, each employing three sensors. Additionally, depth data originating from the system itself is available.

1. INTRODUCTION

Tracking the motion of people has been a topic of active and intense research for the past two decades with many applications. Techniques can be divided into active and passive 2D or 3D tracking ([1, 2, 3]). For a review of the former in the video domain, the reader is referred to [4]. In [5] and [6], a comprehensive review of different passive tracking methods can be found.

3D reconstruction has also received great attention in the past years ([7, 8, 9]). Methods can be either interactive (require manual intervention) or automatic. Two different classes of approaches exist, where different types of information are exploited. The first one includes model-based methods and the second one deals with model-free methods. For reviews of the different methods, the interested reader can be referred to [10] and [11] among others.

In order to facilitate the testing and evaluation of both person tracking algorithms (full-body/body-part/head tracking) and 3D reconstruction algorithms (scene/object reconstruction), test datasets are required. Such datasets that are publicly available are limited in number (e.g. the PETS datasets made available for testing in the last 5 years [12] or [13] for tracking). To fill this void, a number of test recordings were conducted at the I-Lab of the Center for Communications Systems Research in the University of Surrey, UK, using a system of 2 cameras. The outcome was a dataset of visual data, as well as 3D data of the captured scene. Synchronization of the two cameras was also provided, to enable the use of multiple video streams for the same captured scene.

The remainder of the paper is organized as follows. The content of the scenes recorded is described in Section 2. The equipment, acquisition setup and processing of the data are presented in Section 3. A brief description of possible uses of the acquired dataset is provided in Section 4. Finally, the conclusions are drawn in Section 5.

2. SCENE DESCRIPTION

The scenarios of the recorded scenes were selected based on the requirements of video tracking techniques and 3D reconstruction techniques. Most of the scenes were acquired twice, i.e. under optimal lighting conditions (as defined by the studio lights) and sub-optimal lighting conditions created by altering the lighting within the studio, causing generally darker lighting conditions. Sample frames of the recorded sequences are depicted in Figure 1, where each row of images corresponds to the same frame captured by the reference sensor of the reference camera, the reference sensor of the second camera and all three sensors of the reference camera (overlaid) respectively.

The scenes are mostly single subject scenes, where different actions are performed by the subject. More specifically, in some segments of the recorded scenes, the sub-

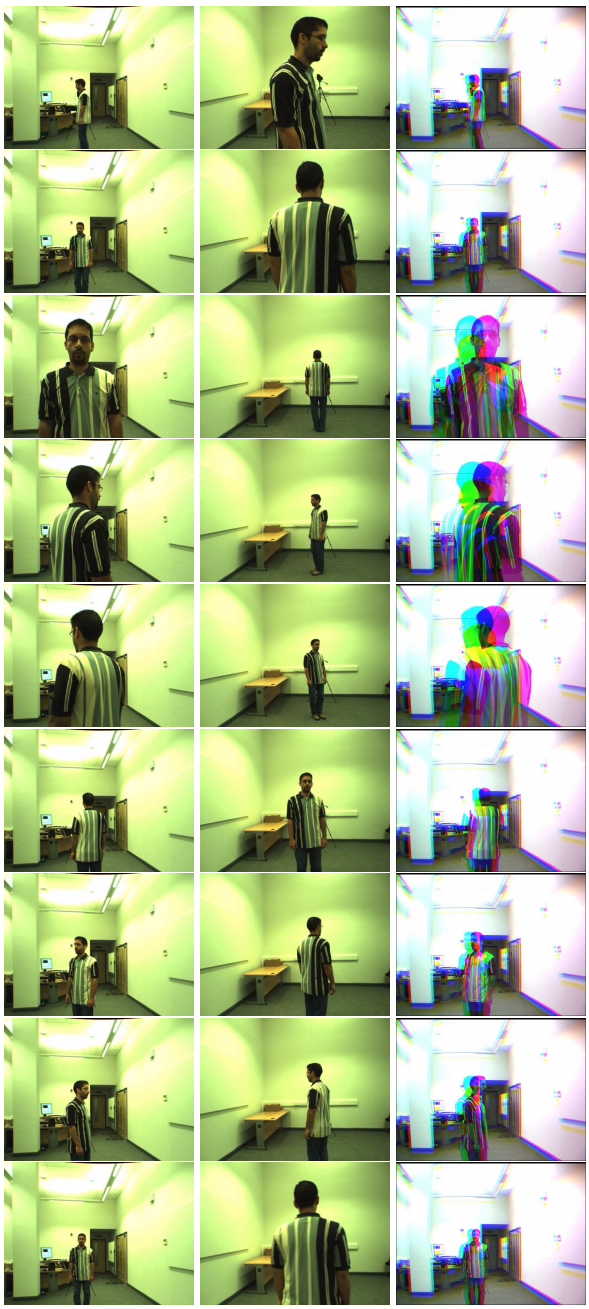

Fig. 1. Sample frames of the recorded sequences. Each row depicts the image captured by the reference sensor of the reference camera, the same image captured by the corresponding sensor of the second camera and all three images captured by the reference camera.

ject is standing still at fixed positions, located at various distances from the camera, with or without occlusion (i.e. self-occlusion). In other segments, the subject is moving on a simple motion trajectory. The latter can be a rectangular path, i.e. in parallel to the camera (left-to-right), forward, parallel again (right-to-left), backwards and so on, or moving along the cameras' axis. In some of the scenes, the person stays within both cameras' field-of-view (FOV) at all times, whereas in others, he moves out of the FOV of at least one camera. Also, in some scenes, the subject is moving in an approximately elliptical path, staying within the FOV at all times. Finally, there exist scene segments where the subject changes the speed of motion, in an effort to assess its impact on tracking algorithms.

Additionally, some scenes were recorded, where two subjects are moving simultaneously, one person is moving while the other is standing at fixed positions (with occlusions taking place) and two subjects are initially standing at different distances from the camera, then moving parallel to the camera and in different directions (with occlusions occurring), then switching positions and moving parallel to the camera again. Finally, there exist scene segments, where the subjects are moving in completely random paths within the cameras FOV (with severe occlusions occurring). Static objects (chairs etc.) are also located within some of the scenes, causing partial occlusions to the subject(s). Required information about the scenes, such as distances from the camera(s), number of subjects, occlusion and lighting conditions etc. have been documented and are also part of the dataset.

3. DATA AND ACQUISITION DESCRIPTION

The video equipment used consisted of a pair of cameras (Digiclops Stereo Vision System), each employing 3 premium progressive scan CCDs capable of capturing in large depth-of-field conditions. A PC was used for monitoring the recording process. The setup of the equipment involved positioning one of the two cameras designated as the reference camera at 87 cm from the back wall, whereas the second camera was positioned so as to face the reference camera at a distance of approximately 4.67 meters. To enable the use of both video streams, a pattern was employed to calibrate both cameras simultaneously. The camera calibration parameters, as well as the transformation data (i.e. details of the position of the second camera with respect to the reference camera) are available along with the dataset. An overview of the setup is illustrated in Figure 2, where the dashed lines indicate the camera's FOV.

The different lighting conditions (i.e. optimal vs. suboptimal) refer to configurations of the studio lights that remained fixed during each scene. The optimal configuration produces light as distributed by the studio lights in their de-

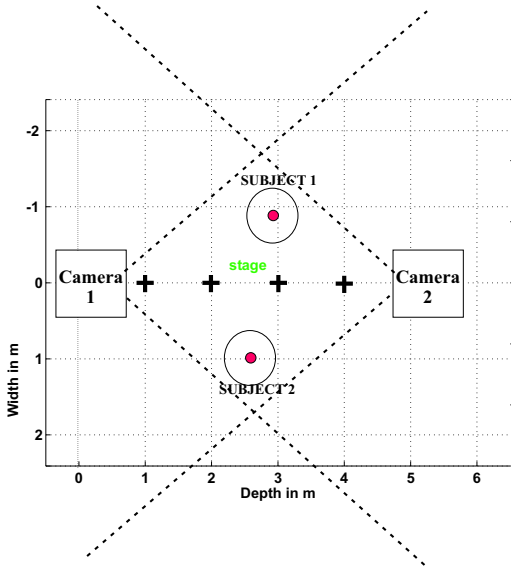

Fig. 2. Recording setup at I-Lab studio.

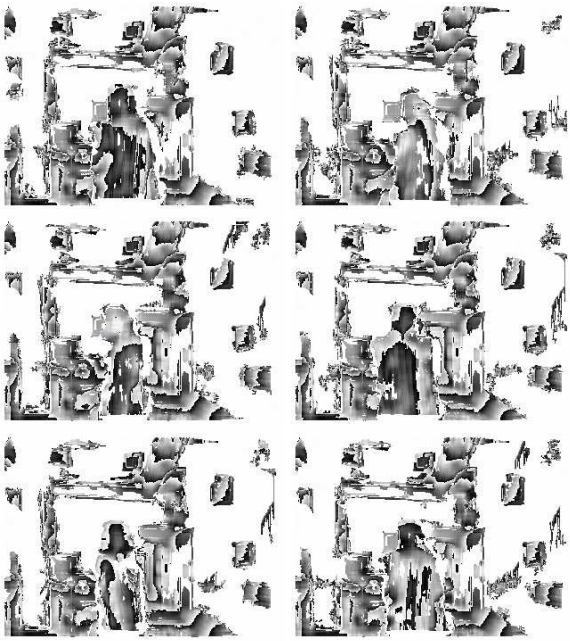

Fig. 3. Sample depth images provided by the reference camera (resolution 320x240).

fault configuration, whereas the sub-optimal configuration causes generally darker lighting conditions.

The scenes were directly recorded on hard disks. Roughly 50 GB of raw data (640x480, 24 bpp, ".ppm" format) was recorded. 6 images are available for each frame of the video sequences, corresponding to the images captured by the three sensors of each camera, as illustrated in the third column of Figure 1. Additionally, depth images and 3D coordinates provided by the system are available. Figure 3 depicts depth images provided by the system itself.

Scene editing involved removing parts that contained no useful information (e.g. segments at the beginning and the end of a take where no action occurred).

4. DESCRIPTION OF POSSIBLE USES

In this section, we will briefly describe possible uses of the scenes recorded. The scene scenarios were mainly selected to enable the testing and evaluation of 3D tracking algorithms. More specifically, segments or entire scenes can be used to perform 3D tracking of faces (heads), body parts, such as arms, limbs and hands (with occlusion). Additionally, 3D reconstruction algorithms can be tested and evaluated using the recorded data originating from 2 (stereo reconstruction) or all three sensors from one camera, or both cameras, in which case full 3D scene or object reconstruction is possible. In both cases, the 3D output of the system can be used as ground truth data for comparison with the 3D reconstruction and tracking algorithms under test. It is worth noting that occlusion scenarios can be tested and evaluated for tracking algorithms, due to the existence of data originating from the extra camera.

Auxiliary uses would include using scene segments where the subject's head performs a full 360 degrees rotation to build a model for the human head. Again the existence of synchronized data from two cameras can help so as to improve the accuracy and quality of the reconstruction. Finally, since both depth information and color information is available, stereoscopic sequence production can be performed (e.g. using off-the-self solutions).

Fig. 4. 3D reconstruction of a single frame captured by the reference camera

In Figure 4, the results of 3D reconstruction of a single frame taken from one of the video sequences captured by the reference camera is illustrated. The algorithm employed uses the Sum of Absolute Differences correlation method to

establish correspondence between the camera images. More specifically, a neighborhood of a given square size is chosen from the reference image (i.e. image captured by the reference sensor of the camera) and is compared against a number of candidate neighborhoods in the other image (along the same row). The best match is selected using the following formula:

$$\min_{d=d_{min}}^{d_{max}} \sum_{i=-\frac{m}{2}}^{i=\frac{m}{2}} \sum_{j=-\frac{m}{2}}^{j=\frac{m}{2}} |I_r[x+i][y+j]-I_l[x+i+d][y+j]| \quad (1)$$

where d_{min} d_{max} are the minimum and maximum disparities, m is the neighborhood (i.e. mask) size and I_r and I_l are the right and left images captured by the camera.

5. CONCLUSION

One of the difficulties involved in the evaluation of the results of computer vision-based algorithms is the absence of test datasets. This work introduced such a dataset for testing 3D reconstruction and 3D person tracking algorithms. The dataset involves a number of scenarios. It consists of video data corresponding to the material recorded, as well as depth data (and 3D position coordinates) of the recorded scene. All data originated from a system of 2 three-sensor cameras. Synchronization of the two cameras is provided. Calibration and transformation parameters are also part of the dataset. A brief description of how the dataset can be used was also provided.

6. ACKNOWLEDGMENT

The work presented was part of integration performed between the Artificial Intelligence and Information Analysis Laboratory, Aristotle University of Thessaloniki, Greece and the Center for Communications Systems Research, University of Surrey, UK, within VISNET, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 programme.

7. REFERENCES

[1] M. La Cascia, S. Sclaroff, and V. Athitsos, "Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3D models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, pp. 322–336, April 2000.

[2] C.S. Wiles, A. Maki, and N. Matsuda, "Hyperpatches for 3D model acquisition and tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 12, pp. 1391–1403, December 2001.

[3] F. Dornaika and J. Ahlberg, "Fast and reliable active appearance model search for 3D face tracking," *IEEE Transactions on Systems, Man and Cybernetics, Part B*, vol. 34, no. 4, pp. 1838–1853, August 2004.

[4] G. Welch and E. Foxlin, "Motion tracking: No silver bullet, but a respectable arsenal," *IEEE Computer Graphics and Applications, special issue on "Tracking"*, vol. 22, no. 6, pp. 24–38, November/December 2002.

[5] J.J. Wang and S. Singh, "Video analysis of human dynamics - a survey," *Real-Time Imaging*, vol. 9, no. 5, pp. 321–346, October 2003.

[6] G. Stamou, M. Krinidis, E. Loutas, N. Nikolaidis, and I. Pitas, "2D and 3D motion tracking in digital video," in *Handbook of Image and Video Processing*, Alan C. Bovik, Ed. Academic Press, 2005.

[7] L. Oisel, E. Memin, L. Morin, and F. Galpin, "One-dimensional dense disparity estimation for three-dimensional reconstruction," *IEEE Transactions on Image Processing*, vol. 12, no. 9, pp. 1107 – 1119, September 2003.

[8] C. Colombo, A. Del Bimbo, and F. Pernici, "Metric 3D reconstruction and texture acquisition of surfaces of revolution from a single uncalibrated view," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 99 – 114, January 2005.

[9] S.F. El-Hakim, E. Whiting, L. Gonzo, and S. Girardi, "3d reconstruction of complex architectures from multiple data," in *3D Virtual Reconstruction and Visualization of Complex Architectures (3D-Arch2005)*, Venice-Mestre, Italy, August 2005.

[10] G. Slabaugh, W.B. Culbertson, T. Malzbender, and R. Schafer, "Survey of volumetric scene reconstruction methods from photographs," in *Joint IEEE TCVG and Eurographics Workshop*, Stony Brook, New York, USA, June 2001, pp. 81–100.

[11] F. Bernardini and H. Rushmeier, "The 3D model acquisition pipeline," *Computer Graphics Forum*, vol. 21, no. 2, pp. 149–172, 2002.

[12] "PETS dataset repository for testing tracking and surveillance algorithms," `ftp://ftp.pets.rdg.ac.uk`.

[13] M. Krinidis, G. Stamou, H. Teutsch, S. Spors, N. Nikolaidis, R. Rabenstein, and I. Pitas, "An audio-visual database for evaluating person tracking algorithms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005)*, Philadelphia, USA, March 2005.