AN AUDIO-VISUAL DATABASE FOR EVALUATING PERSON TRACKING ALGORITHMS

M. Krinidis[†]*, G. Stamou*[†]*, H. Teutsch*[‡]*, S. Spors*[‡]*, N. Nikolaidis*[†]*, R. Rabenstein*[‡] *and I. Pitas*[†]

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

[‡] Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg, Cauerstrasse 7, 91058 Erlangen, GERMANY
Email: {*teutsch,spors,rabe*}*@nt.e-technik.uni-erlangen.de*

ABSTRACT

This paper presents an audio-visual database that can be used as a reference database for testing and evaluation of video, audio or joint audio-visual person tracking algorithms, as well as speaker localization methods. Additional possible uses include the testing of face detection and pose estimation algorithms. A number of different scenes are included in the database, ranging from simple to complex scenes that can challenge existing algorithms. They include different subjects, with appearances that can cause problems to video tracking algorithms, (e.g. facial features such as beards, glasses, etc.), optimal and artificially created sub-optimal lighting conditions, subject movement based on simple as well as random motion trajectories, different distances from the camera/microphones and occlusion. The database incorporates ground truth data (3-D position in time) originating from a commercially available 4-camera infrared (IR) tracking system. Examples of how the database can be used to evaluate video and audio tracking algorithms are also provided.

1. INTRODUCTION

Tracking the motion of people has been a topic of active and intense research for the past two decades. Techniques can be divided into active and passive tracking. For a review of the former in the video domain, the reader is referred to [1]. Passive tracking techniques have exploited data from different modalities, such as audio [2], video [1], infrared [3], range data etc., in an effort to provide accurate estimation results. Systems that employ a scheme for fusing data from more than one modality have also been proposed [4]. In order to facilitate the evaluation of passive person tracking algorithms, reference databases that include scenes with different subjects, lighting conditions, motion trajectories and occlusion, multiple simultaneously active acoustic sources, as well as ground truth data, are required. To the best of the authors' knowledge, such reference databases that are publicly available and cover a wide range of tracking scenarios do not exist. To fill this void, a number of test recordings were conducted at the Virtual Studio (VS) of the Institute for Media Technologies in the Technical University of Ilmenau, Germany. The outcome was a database of audiovisual data, as well as ground truth data, i.e. the 3-D position coordinates of the subject(s). Synchronization of audio and video data was also provided, to enable its use for the evaluation of joint audio-visual tracking algorithms. More specifically, time stamps were generated, distributed to the video and audio recording equipment and recorded along with the audio/video data.

The remainder of the paper is organized as follows. The different scenes recorded are described in Section 2. The equipment, acquisition setup and post-processing of the video and audio data are presented in Sections 3 and 4 respectively, while the ground truth data are described in Section 5. Brief examples of how video-based and audio-based tracking algorithms can be evaluated with this database are presented in Section 6. Finally, the conclusions are drawn in Section 7.

2. SCENE DESCRIPTIONS

Based on the requirements of video tracking techniques the scene conditions (e.g. lighting conditions, motion trajectories of the subjects, occlusion etc.) have been selected such that they constitute a range of increasingly complex tracking scenarios. They attempt to cover as many failure modes of current audio/video tracking algorithms as possible. For example, scenes where the subjects' appearance (beards, glasses, clothing with skin-like color etc.) or the trajectory of the movement could cause problems to video-based tracking algorithms were recorded. Most of the scenes were acquired twice, i.e. under optimal lighting conditions (as defined by the studio technicians) and sub-optimal lighting conditions created using the lighting equipment of the studio, thereby introducing hard shadows and bright/dark areas in the recorded video sequences. Sample frames of the recorded sequences illustrating the number of subjects, different lighting and occlusion conditions, motion trajectories etc. are depicted in Figure 1.

To facilitate the evaluation of acoustic tracking techniques, single as well as multiple simultaneously active subjects that were either fixed or moving were recorded in the fairly reverberant environment of the VS. A "teleprompter", i.e. an electronic television prompting device was used.

The scenes can be loosely categorized according to the number of subjects. The first category consists of a number of scenes that involve a single person. Such scenes include cases where the subject is standing still at fixed positions, located at various distances from the camera, while speaking and exhibiting limited head and hands motion, with or without occlusion (i.e. self-occlusion). Additional scenes, where the subject is moving on a simple motion trajectory, are also available. In several scenes, the movement

Scene No.	Subjects	Fixed/Moving	Distance(m)	Lighting	Occlusion	Motion	Timecode
1	1	Fixed	2	Optimal	No	0	10:04:34:00
12	1	Fixed	6	Sub-optimal	Yes	0	10:16:39:20
20	1	Moving	4	Sub-optimal	Yes	2	10:25:42:24
25	2	Moving, Fixed	4, 6	Sub-optimal	Yes	4	10:31:40:17
35	2	Moving	–	Optimal	Yes	4	10:45:53:10
40	5	Moving	–	Sub-optimal	Yes	4	10:54:40:24
43	1	Moving	2	Optimal	No	6	11:02:05:02
54	1	Moving (fast)	–	Optimal	No	3	11:14:53:24

Table 1. Excerpt of the scene documentation (see text for details)

occurs on a rectangular path, i.e. in parallel to the camera (left-to-right), forward, parallel again (right-to-left), backwards and so on. In some of the scenes, the person stays within the camera field-of-view (FOV) at all times, whereas in others, he moves out of the FOV to provide additional test conditions, e.g. testing of the initialization/re-initialization of a fully automatic video-based tracking system. Single-subject scenes also include takes where the subject is moving in an approximately elliptical path, staying within the FOV at all times. Finally, there exist scenes where the main goal is to assess the impact of the subject's speed of motion. In these scenes, the person is moving in parallel to the camera (left to right), or towards the camera and back, at different speeds, while talking naturally.

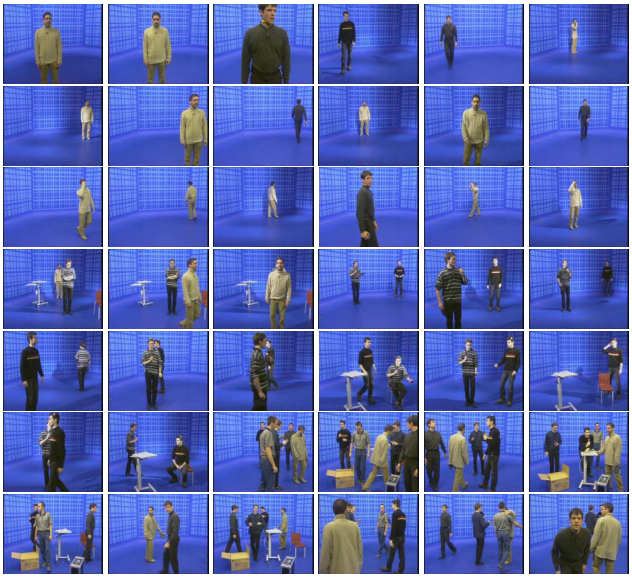

Fig. 1. Sample frames of the recorded sequences.

The second category involves scenes with two subjects which, at times, were talking simultaneously. In some of them, the subjects are standing at different fixed positions, whereas in others, one person is moving while the other is standing at fixed positions (with occlusions taking place). Additional scenes shot involve two people 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. This category concludes with "life-like" scenes, where the subjects are moving in completely random paths within the camera FOV (with severe occlusions occurring). The last category also consists of "life-like" scenes, but more than two people (up to five) are moving in random patterns. Static objects (tables, chairs etc.) are also located within the scene, causing partial occlusions to the subjects.

A total of 55 scenes are available in the database. Required information about the scenes, such as distances from the camera, number of subjects, timecode, occlusion and lighting conditions etc. have been documented and are also part of the database. A sample of the documentation is illustrated in Table 1. Note that motion trajectories have been coded (e.g. 4 in the motion column corresponds to random movement).

3. VIDEO DATA

The video equipment used consisted of a studio quality video camera (JVC-KY19) with a monitoring unit and a BTS D1 digital video recorder. Several Silicon Graphics workstations were used for monitoring the recording process and for the coordination of the infra-red tracking system. The setup of the equipment involved positioning the camera at a distance of approximately 6.5 meters from the rear panel of the studio, at a height of approximately 1.5 meters and tilted by 30 degrees in the vertical axis, facing downwards. To enable the use of the database for 3-D tracking, a planar chessboard pattern was employed to calibrate the camera using the method described in [5]. The camera calibration parameters, as well as the reference coordinate systems (video, audio, infra-red) are available as part of the database. In order for the actors to have visual aids during the recording, several points were marked on the studio floor, indicating different distances from the camera. An overview of the setup is illustrated in Figure 2, where the dashed lines indicate the camera's FOV. Visual aids are denoted by (+).

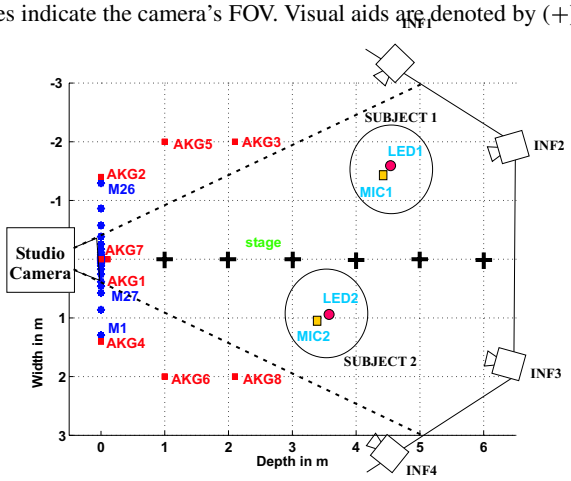

Fig. 2. Recording setup at the Virtual Studio.

The background of the recorded video sequences is illustrated in Figure 1. The different lighting conditions (i.e. optimal vs. sub-optimal) refer to configurations of the studio lights that remained fixed during each scene. The optimal configuration produces light distributed uniformly in the stage of the studio causing soft shadows, whereas the sub-optimal configuration introduces hard shadows and causes bright and dark areas to appear in the recorded video sequences, thereby challenging video-based person tracking algorithms.

The scenes were originally recorded on D1 digital tapes. The material was then extracted, resulting in roughly 100 GB of raw video data in full PAL resolution (25fps, 4:2:2, 720x576, 24 bpp) and then converted to the popular AVI format (uncompressed) to enable universal processing. Some interlacing problems appear in scenes with significant motion. 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). The above procedure resulted in approximately 40 minutes of processed, usable recording material.

4. AUDIO DATA

In order to assess the performance of both sound recording by means of digital beamforming arrays and acoustic tracking, a total of 37 microphone channels were recorded. The microphone arrangement consists of a linear microphone array, separate microphones surrounding the stage, and two close-talking microphones, as depicted in Figure 2. The 26 stars, labeled M1...M26 denote the omnidirectional electret capsules (Panasonic WM-60A) that are used to form a wide-bandwidth linear microphone array designed for high-quality sound acquisition (see [6]). M27 is an additional electret capsule that can be used for three-dimensional acoustic source tracking in combination with at least three microphones from the linear array. All 27 channels (M1...M27) were recorded into multi-channel wave-files synchronously using a sampling rate of 48 kHz and a resolution of 24 bits/sample. The hardware and software utilized for these recordings were designed at the University of Erlangen-Nuremberg. Eight additional AKG CK92 omnidirectional condenser microphones, shown as boxes labeled AKG1...AKG8 in Figure 2, were placed in front of and along the sides of the stage. Moreover, a maximum of two close-talking microphones, depending on whether one or two acoustic sources were present on stage, were recorded (denoted by MIC1 and MIC2 in Figure 2). Again, these nine/ten microphone signals were recorded synchronously using a sampling rate of 48 kHz and a resolution of 24 bits/sample utilizing a ProTools system. In total, the database contains approximately 20 GB of audio data.

The reverberation time for an acoustic source approximately 2 m in front of the camera has been found to be in the order of 700 ms. At a first glance this seems to be resulting in an extremely challenging acoustic environment for any acoustic source localization algorithm trying to estimate source positions reliably. One has to keep in mind, though, that the VS is built into a fairly large hall (12 m $\times$ 20 m). Many acoustic source localization algorithms process the data in a block-based fashion, where the blocklength is typically in the order of 30 ms, thereby, to some extent, minimizing the effect of late reverberation.

5. GROUND TRUTH DATA

To enable the use of the database by joint audio-visual person tracking or speaker identification and tracking algorithms, synchronization of the audio and video tracks is necessary. For that purpose, time stamps have been generated by the Virtual Studio according to the SMPTE time code standard. The time code was distributed to the video recorder and the ProTools audio recording system and recorded along with the audio/video data.

To objectively evaluate the results of any person tracking algorithm, ground truth data which includes the actual position of the subjects in the recorded scenes are required. For this purpose, a 4-camera infrared tracking system [3] has been used, which tracks the position of 2 mobile, battery operated, infrared emitters usually placed on the actors' heads, as seen in Figure 2, where INF1 ... INF4 denote the infrared cameras and LED1, LED2 denote the mobile emitters. Note that only two IR emitters were in use even if the number of subjects was greater than two. The system records the subjects' 3-D position. It produces position results that are sufficiently accurate for the evaluation of purely audiovisual tracking methods without any active or passive markers.

6. TRACKING ALGORITHMS EVALUATION EXAMPLES

In this section, we will briefly present examples of how the database can be used to evaluate video-based and audio-based tracking algorithms. Therefore, the usefulness of the available ground truth data in performance evaluation of tracking algorithms will be illustrated.

Monocular video-based tracking is extremely challenging. Furthermore, 3-D tracking often requires the use of a calibrated camera and additional constraints. A number of algorithms have been tested on the database, e.g. [7, 8].

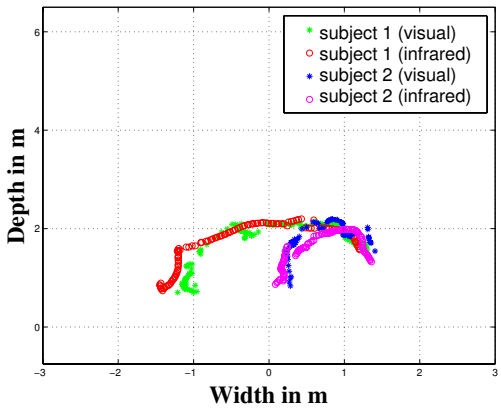

Fig. 3. Performance of a video-based tracking algorithm

In Figure 3, the results of an automatic face detection and tracking system, taken from scene 35 of the database, are illustrated. The tracking algorithm selects a large number of point features in the tracking region which are subsequently tracked in the next frames. Initialization can be performed either manually or by means of an automatic face detection module, based on color [7] and Haar-features [9]. Feature generation is based on an algorithm used for point feature tracking [10]. Point features are tracked using a modified variant of the Kanade-Lucas-Tomasi (KLT) algorithm [10]. Calculation of the 3-D (world) coordinates of the tracked object(s) is possible, since a calibrated camera has been used. The results in Fig. 3 correspond to the X and Z coordinates of the subjects. Stars correspond to the results of the video tracking

algorithm, while circles denote the output of the infrared tracking system.

Acoustic source localization algorithms try to extract the three-dimensional positions of possibly moving sound sources by appropriately processing signals provided by typically four spatially separated microphones. The audio database has been used to test a real-time capable implementation of various localization algorithms. For details concerning the implementation, the interested reader is referred to [6]. The majority of the algorithms implemented are described in detail in [2] and [11].

Figure 4 shows the performance of an acoustic tracking system utilizing a GCC-based (Generalized Cross-Correlation) algorithm on the signals recorded during scene 43 of the database with microphones M1, M12, M26, and M27. The stars show the output of the source tracking algorithm and the circles denote the output of the infrared tracking system. Note that the performance in only two dimensions is shown, although both the LED-based tracking system as well as the acoustic source tracking system are capable of three-dimensional tracking.

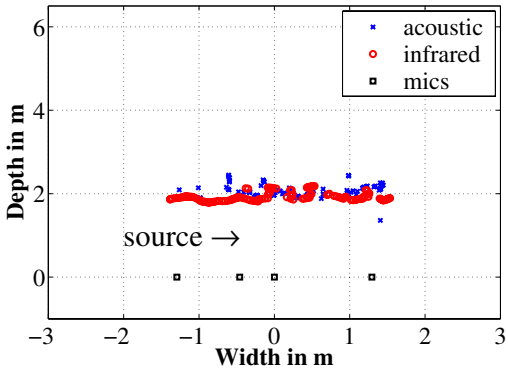

Fig. 4. Performance of an audio-based tracking algorithm

The database described in this paper has been created to also enable the further development of joint audio-visual tracking methods. Audio-visual tracking applies data fusion techniques to combine the estimates from mono-modal localizers based on only either video cameras or microphone arrays. Combining information from video and audio recordings has the potential for more robust tracking estimates. For a survey of the literature and description of audio-visual data fusion by recursive state estimation, see e.g. [4, 2].

7. CONCLUSION

One of the difficulties involved in the evaluation of the results of person tracking systems is the absence of test databases. This paper introduced an audio-visual reference database, for testing audio, video and joint audiovisual person tracking algorithms. The database involves simple and challenging human tracking scenarios. It consists of video and audio data corresponding to the recorded material, as well as ground truth data (3-D position coordinates) of the subject(s), originating from a 4-camera infrared (IR) tracking system. Synchronization of the audio and video tracks is also provided. Calibration parameters, as well as reference coordinate systems (audio, video, infrared) are also part of the database. Examples of how the database can be used for evaluation of person tracking algorithms were briefly described. The database is freely available for research purposes at *http://www.aiia.csd.auth.gr*.

8. ACKNOWLEDGMENT

The database described above has been recorded within the framework of the project CARROUSO, IST-1999 20993, while postprocessing of the video data has been performed within the framework of the SIMILAR European Network of Excellence on multimodal interfaces (www.similar.cc), both funded by the IST Programme of the Commission of the European Communities. The authors express their sincere thanks to the Institute for Media Technologies of the Technical University of Ilmenau, Germany, for the use of their Virtual Studio.

9. REFERENCES

[1] 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.

[2] M.S. Brandstein and D.B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer Verlag, 2001.

[3] "The ORAD led-based tracking system," `http://www.orad.co.il`.

[4] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video object localization and tracking," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 22–31, January 2001.

[5] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Seventh IEEE International Conference on Computer Vision (ICCV99)*, Corfu, Greece, September 1999, vol. 1, pp. 667–673.

[6] H. Teutsch, S. Spors, W. Herbordt, W. Kellermann, and R. Rabenstein, "An integrated real-time system for immersive audio applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2003.

[7] K. Sobottka and I. Pitas, "Looking for faces and facial features in color images," *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, Russian Academy of Sciences*, vol. 7, no. 1, pp. 124–137, 1997.

[8] E. Loutas, N. Nikolaidis, and I. Pitas, "A mutual information approach to articulated object tracking," in *International Symposium on Circuits and Systems (ISCAS '03)*, Bangkok - Thailand, May 2003, vol. 2, pp. 672–675.

[9] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *IEEE International Conference on Image Processing (ICIP02)*, Rochester, New York, USA, September 2002, pp. 900–903.

[10] J. Shi and C. Tomasi, "Good features to track.," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR94)*, Seattle, United States, June 1994, pp. 593–600.

[11] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Am.*, vol. 107, no. 1, pp. 384–391, January 2000.