FEATURE-BASED TRACKING ON A MULTI-OMNIDIRECTIONAL CAMERA DATASET

Barış Evrim Demiröz, İsmail Arı, Orhan Eroğlu, Albert Ali Salah, Lale Akarun

Boğaziçi University, Computer Engineering Department
{baris.demiroz, ismailar, orhan.eroglu, salah, akarun}@boun.edu.tr

ABSTRACT

Omnidirectional cameras have a lot of potential for surveillance and ambient intelligence applications, since they provide increased coverage with fewer cameras. We introduce the new BOMNI dataset, collected with two omnidirectional cameras simultaneously. The dataset contains single subject and multi-subject interaction scenarios, as well as actions relevant for ambient assisted living, such as falling down. We describe evaluation protocols on this dataset, and provide benchmarking baseline results for two tracking systems based on bounding box and interest point matching after foreground-background segmentation, respectively.

Index Terms— Video surveillance, Object tracking, Omnidirectional cameras, Image sequence analysis, AAL dataset

1. INTRODUCTION

Monitoring a wide area with conventional directional cameras requires the usage of either multiple cameras or Pan-Tilt-Zoom(PTZ) cameras. In this paper, we consider the usage of omnidirectional cameras for indoor monitoring, which have advantages over conventional cameras.

Using multiple cameras results in increased cost of hardware and maintenance, compared to a single camera setup. Furthermore, bandwidth requirements increase with each additional camera, and there will be synchronization issues and maintenance of multiple images acquired from each camera. The main advantage of a system using multiple directional cameras is the ability to obtain high resolution images from the monitored directions. Although PTZ cameras are advantageous in terms of bandwidth utilization and cost, they are able to capture an image from only one direction at any given instant. In addition to this, the presence of moving mechanical parts makes them prone to failures. By sacrificing from resolution, the image of the whole scene can be obtained continuously using one omnidirectional camera. These properties of omnidirectional cameras make them attractive for surveillance type applications.

In this paper we introduce a novel database for tracking (and monitoring) people in an indoor environment with omnidirectional cameras. A good omnidirectional video database for tracking must include several conditions and challenges taken from the real-world problems to be successful. In a real indoor setup, one can easily observe that motions of persons may be complex, and they may completely or partially disappear because of occlusions by other objects, other persons or even by their own body parts. Due to windows and lighting changes, illumination varies through time, and therefore, objects may have different appearances across the recordings. An action may be performed by different individuals with various speeds and trajectories. Noisy frames and complex backgrounds are further major challenges in real world videos. The database we describe here contains all these variations.

This paper is structured as follows. Section 2 briefly discusses the related benchmarking work in the domain of human behavior analysis via omnidirectional cameras. Section 3 details the setup of the collected database, the variation of conditions and scenarios, and gives related technical details. Section 4 describes benchmark tracking algorithms, and reports performance under widely used evaluation criteria. Finally, Section 5 concludes the paper.

2. RELATED WORK

Omnidirectional cameras come in different flavors. There are systems that construct the omnidirectional field by stitching multiple captured images, or by taking a single image with a special lens (e.g. fish-eye) or a lens-mirror combination (catadioptric devices). Depending on the sensor, the acquired images show different image geometry conditions. In this section, we shortly review related work in the field of surveillance and human motion analysis performed by data acquired from omnidirectional sensors. The work using conventional cameras is extensively reviewed in the literature.

A number of databases are publicly available for surveillance systems. Most such databases are collected using conventional directional cameras. Typically, resolutions are between 640×480 and 768×576, with frame rates of 25fps or less.

Table 1 compares different databases that use omnidirectional cameras and summarizes their main characteristics. The PETS 2001 Database 4 contains an omnidirectional and a moving conventional camera, and its purpose is vehicle and people tracking in an outdoor scenario. It contains about 6800

This work is supported by Bogazici University research project BAP-6531.

Databases	Environment	Resolution	Fps	Number of Cameras
PETS 2001	Outdoor	768×576	25	1+1
BOSS	Indoor	720×576	25	9
PETS-ICVS	Indoor	720×576	25	1 + 2
AMI Meeting	Indoor	720×576	25	2+4
CLEAR	Indoor	see text	15-30	1+5
TAU-DANCE	Indoor	768×576	25	1+4
BOMNI (this work)	Indoor	640×480	7	2

Table 1: Comparison of different omnidirectional video databases.

frames for training and 5000 frames for testing[1]. All the other databases we list are collected from indoor scenarios. The AMI Meeting corpus is not primarily an omnidirectional video database, but it contains (in addition to four conventional cameras) two semi-fisheye lens cameras, and records meetings of several people sitting around a table or presenting something on a whiteboard [1]. It is not very suitable for testing tracking applications, as the subjects are not moving around too much. The PETS-ICVS database similarly contains an around-the-table scenario, where there is one omnidirectional camera on top of the table, and two conventional cameras directed to the table to capture facial expressions[2]. The considered actions also involve typical meeting actions like talking, raising hands, nodding, getting up, yawning, etc. A more extensive but similar scenario is followed in the CHIL project and the subsequent CLEAR evaluation campaigns [2], where an omnidirectional camera at the ceiling is complemented by five conventional cameras. CHIL duplicated this setup over five locations, and recording conditions varied depending on location and session (640×480 to 1024×768 resolution and 15-30 fps).

The Tel Aviv Univ. Dance database uses an omnidirectional camera in addition to four conventional cameras to record several types of dance movements performed by different subjects [3]. This database does not have occlusions or illumination changes. Finally, the BOSS database is an indoor setup and constructed for the needs of passenger security and remote diagnostics, or predictive maintenance in public transportation vehicles [3]. With high frame rates and resolution, it consists of 15 sequences in which actions such as cell phone theft, disease (fainting) and harassment may be observed.

Acquiring images from multiple cameras improves the coverage and helps in dealing with occlusions. The omnidirectional video databases we mention in this section are all acquired with multiple cameras, but none of them use multiple omnidirectional cameras. In meeting corpora, the omnidirectional ceiling camera has been found to be an adequate modality to do the manual annotations, whereas conventional cameras are used to track people [4]. There are also hybrid systems that combine omnidirectional cameras with for instance PTZ cameras [5].

For the specific application of tracking people from omnidirectional camera the proposed techniques are largely similar to tracking in conventional cameras, Boult et al. proposed a system for tracking camouflaged targets using an omnidirectional camera, which is based on background modeling followed by connected component analysis [6]. In another study, Wang et al. utilized the CamShift algorithm and optical flow to track moving objects in the scene [7]. In addition to tracking, fall detection is also performed using calibrated omnidirectional camera in [7]. The CLEAR dataset was used in [8] to implement a hybrid probabilistic neural modal for person tracking with good results.

In the next section, we describe the BOMNI database.

3. THE BOMNI DATABASE

3.1. Setup

Our database includes samples taken by two omnidirectional cameras. The first camera is mounted on the ceiling of the room and the latter is fixed on a side wall. The room is almost square-shaped, approximately $7m \times 7m$. Although there are many objects cluttering the room, only two chairs, two tables and a sink are actively used by the subjects. A rough plan of the room and the location of interacted objects can be seen in Figure 1.

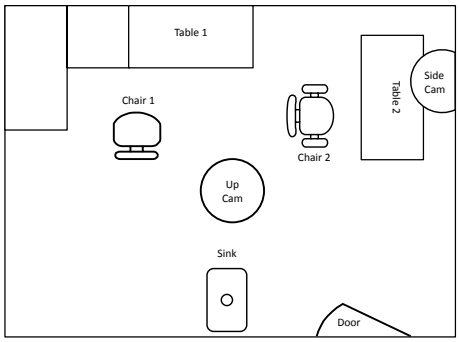

Fig. 1: Floor plan of the room that is used for the data acquisition.

Videos have been acquired using Oncam IPC, which is a 360° ceiling-mounted IP camera with a 5 megapixel sensor with no moving parts [4]. Calibrated camera parameters are distributed along with the dataset. Although the camera supports higher resolutions, the videos were captured using 640×480 resolution because of network bandwidth limitations. For the

[1]PETS2001 Dataset, http://www.cvg.cs.rdg.ac.uk/PETS2001/pets2001-dataset.html

[2]PETS-ICVS Dataset, http://www.cvg.cs.rdg.ac.uk/PETS-ICVS/pets-icvs-db.html

[3]BOSS On Board Wireless Secured Video Surveillance dataset,http://www.multitel.be/image/research-development/research-projects/boss.php

[4]Oncam IPC Conceal-mounting IP Camera, http://www.oncamglobal.net/31-oncam-ipc.html

same reason, the frame rate of the original videos are around 8 fps and there are occasional frame drops. Cameras are also configured to work in auto exposure and white balance mode resulting in infrequent intensity changes for the whole captured frames.

The only light source is the natural light entering room through windows located at one side of the room. At the time of video acquisition the light source is not controlled, allowing the illumination to change between videos. To illustrate the illumination variation, the standard deviation of image pixels' intensity values are calculated using the first 30 frames of videos, and in HSV color space. The pooled standard deviation for the V channel is 14.99, and 11.73 for top and side view, respectively (See Figure 2). Furthermore, from the viewpoint of the side camera, the subjects are occluded by the objects in the room while entering and exiting the room, and in the multi-user scenario there are subjects occluding each other. Low frame rate, frame drops, occlusions and uncontrolled illumination are the usual conditions faced in a real world application, and form the challenges posed by the dataset.

Fig. 2: Illustration of the standard deviation of image pixel intensities for the V channel of HSV color space. Darker colors indicate higher deviation.

3.2. Acquisition Scenarios

We recorded data from two different acquisition scenarios, single and multi-user, respectively. Scenario 1 consists of recordings of five subjects from both cameras resulting in a total of 10 videos. In each video, a subject enters to the room, walks across to the table, grabs a bottle, sits down on chair 1, drinks water, stands up, puts the bottle back on to the table, goes to the sink, washes hands, exits the room, enters to the room after a short delay, walks to chair 2, sits down, idles for a short time, stands up, walks across the room to table 1, grabs a bowl, while walking across the room faints and falls down; in this order. In this scenario, there are only minor and rare occlusions caused by the chairs. Sample frames of Scenario 1 can be seen in Figure 3.

Scenario 2 presents videos of multiple people interacting with each other. In each video, the actions taking place are as follows: the first person (A) enters to the room, walks across the room, sits down on chair 1, the second person (B) enters to the room, walks across the room to table 1, A stands up, A and

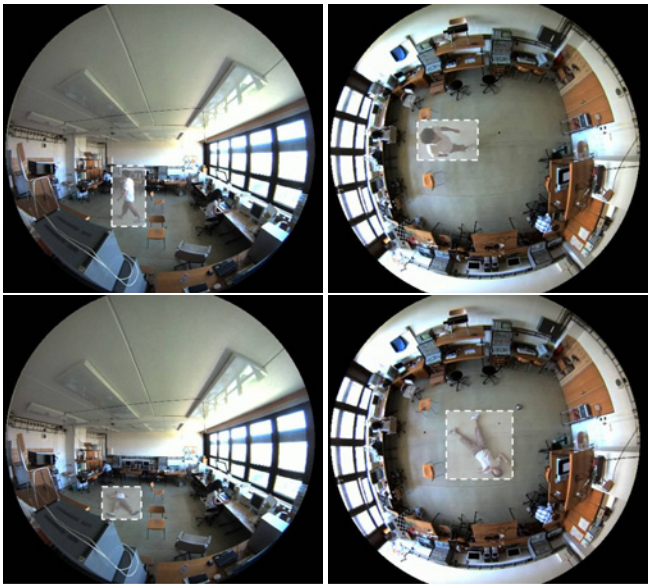

Fig. 3: Sample frames of Scenario 1. Annotations are depicted in lighter color for clarity.

B shake hands, the third person (C) enters the room, A and C meet and shake hands in the middle of the room, B starts walking and exits room, A and C walk to table 2, A and C look at an object for a short duration, A starts walking towards the door, C starts walking, A exits the room, C exits the room. The subjects' appearances on the videos, especially the ones taken with the side camera, suffer from serious occlusions during the performance of the scenario. Five subjects in total formed three groups of size three, and for each group, by changing roles in group, six videos are recorded. As a result, 36 videos are acquired from both cameras. Samples of Scenario 2 can be seen in Figure 4.

All videos have been recorded in MJPEG format and annotated using interactive video annotation tool *vatic* [9]. Database annotation contains the bounding box of the subjects performing various actions. There are six subjects in total, one of whom is female.

The database also contains hand-made annotations for action segments. In the single person scenario, six actions are annotated. These are sitting, walking, drinking, washing-hands, fainting (laying on the floor) and the opening-closing-door actions, respectively. The three person scenario contains five actions, which are slightly different. Fainting for instance is not relevant for a multi-person scenario, but interactive actions are considered. In particular, we have annotated sitting, walking, standing, shaking-hands and interested-in-object actions, respectively. The number of annotated actions for each case and the total number of frames for each action are given together with annotation files on the database webpage.

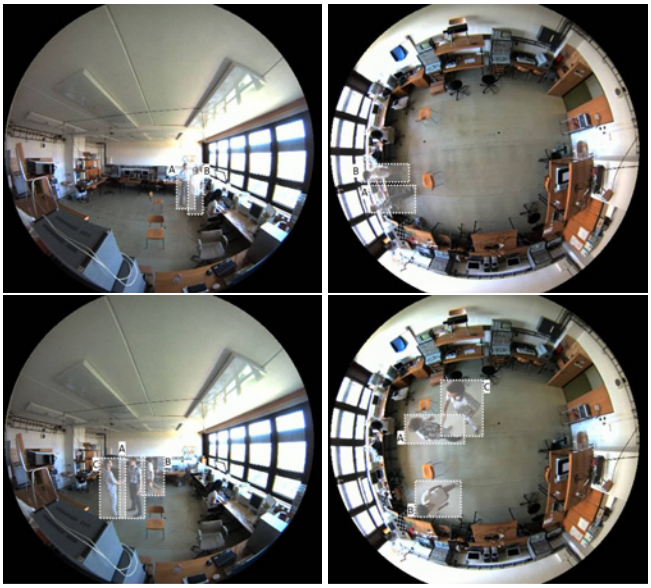

Fig. 4: Sample frames of Scenario 2. Annotations are depicted in lighter color for clarity.

4. TRACKING BASELINE

Comparison of different trackers on a dataset needs common performance measures. For the proposed dataset, we advise to use the CLEAR Multiple Object Tracking metrics [2], MOTP and MOTA, respectively.

Multiple object tracking precision (MOTP) is defined as:

$$\text{MOTP} = \frac{\sum_{i,k,t} d^i_{k,t}}{\sum_{k,t} c_{k,t}} \tag{1}$$

where $d^i_{k,t}$ is the distance between the object o_i and its corresponding hypothesis for frame t of the k^{th} video, $c_{k,t}$ is the number of matches found for the frame t of k^{th} video.

Multiple object tracking accuracy (MOTA) is defined as:

$$\text{MOTA} = 1 - \frac{\sum_{k,t}(m_{k,t} + fp_{k,t} + mme_{k,t})}{\sum_{k,t} g_{k,t}} \tag{2}$$

where $m_{k,t}$, $fp_{k,t}$ and $mme_{k,t}$ are the number of misses, false positives and mismatches, respectively, for the frame t of k^{th} video. The distance is defined as a function of two regions:

$$\text{distance}(o, h) = 1 - \frac{o \cap h}{o \cup h} \tag{3}$$

with o denoting the object, and h the hypothesis, respectively. A threshold on this distance is selected as 0.5 to reject a match between a hypothesis and an object.

We provide a baseline tracker that utilizes foreground detection. The algorithm consists of two main parts: the first part models the background using mixture of Gaussians, and the second part tracks the blobs marked as foreground by matching blobs between consecutive frames.

For foreground segmentation, intensity values of each pixel are modeled as a mixture of Gaussians. The model is learned from the first few frames of the video and updated as the frames continue to arrive [10]. Shadows of the foreground objects are usually marked as foreground using this approach. To overcome this problem, a specific amount of chromatic distortion and intensity distortion from the background model is allowed. The number of Gaussians are selected adaptively, to avoid overfitting and underfitting [11].

After obtaining foreground regions, morphological closing is applied to remove discontinuities between blobs that belong to the same region. Small blobs are unlikely to be foreground objects, so they are removed from our foreground candidates and each of the resulting blobs are considered to represent an object (see Figure 5).

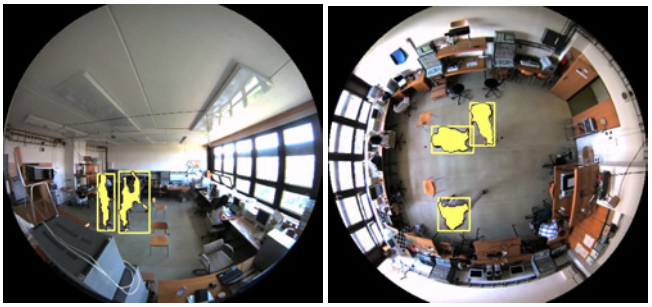

Fig. 5: Left: Merged subjects due to occlusion, Right: Successful subject segmentation.

Two different matchers are tested. For *bounding box based matching, distances are calculated between bounding boxes of the current frame and the previous frame. Then, labels are assigned using the Hungarian algorithm. If any regions are left unlabeled, a new label is assigned to that region.*

The second tested scheme involves interest point based matching, which are detected inside the bounding box of each blob using FAST corner detection [12]. Using bounding rectangles instead of the actual segmentation provided better results, since foreground detection can result in smaller regions than the actual foreground. BRIEF (Binary Robust Independent Elementary Features) descriptors are extracted from each interest point [13]. BRIEF is originally designed to work on grayscale images, but we extended it to work on color images to improve the descriptor power [14]. We followed the color boosting transformation in the opponent color space. 256 bits are used for each channel and concatenated to form a final feature vector involving 96 bytes.

For a given frame, for each interest point of each blob, the closest interest point with Hamming distance lower than a threshold is searched in previous n frames. It is assumed that an interest point descriptor does not change significantly between n consequent frames. For that reason n should be

small. A counter for the label of the corresponding blob of an interest point is incremented. Finally, the label assignments are done using the Hungarian algorithm. If any blob is left unlabeled, a new label is created and assigned to that blob.

Algorithm	MOTP (overlap)	Mismatch
Bounding Box Based	0.73	0.49%
Interest Point Based	0.73	0.44%

Table 2: Proposed algorithm evaluation according to the CLEAR metrics and comparison of blob matchers.

The results of the proposed algorithms can be seen in Table 2 and Table 3. Although the precisions of the algorithms are comparable, using appearance information in matching step improves mismatch error by 8%. Miss and false positive ratios are omitted in Table 2 since they are determined by the foreground segmentation step and are the same for both matchers. Our interest point based algorithm has slightly better precision for the side view, but the heavy occlusions reduce accuracy dramatically.

Videos	MOTP (overlap)	MOTA	Miss	False Positive	Mismatch
All	0.73	68.18%	23.62%	7.74%	0.44%
Top view	0.72	73.52%	18.78%	7.35%	0.32%
Side view	0.74	62.21%	28.95%	8.17%	0.66%

Table 3: Comparison of the tracking results using interest point based matcher with respect to camera position.

5. CONCLUSIONS

We have described the first indoor multi-omnidirectional camera dataset for activity recognition and provided benchmark algorithms for tracking. The database, its subject annotations, timing of specific actions of the subjects and a Java implementation of performance evaluation are available through http://bit.ly/BOMNI-DB. We focus on tracking in this paper and leave action recognition for future work. We believe this open multi-camera dataset will be a useful contribution to the community for indoors surveillance and ambient assisted living applications.

Acknowledgments

This database has been collected during the eNTERFACE'11 Workshop on Multimodal Interfaces. We would like to thank to Milos Zelezny and to others from University of West Bohemia for organizing the workshop. We would also like to thank to Alexey Karpov and Alexander Ronzhin for their collaboration. We would like to thank all subjects who performed in videos for us. We are grateful to Carl Vondrick for the great *vatic* tool and for his quick feedback on problems we have faced. We are very thankful to Hakan Anıt, Aysun Çoban, Serhat Mercan from Boğaziçi University, Mustafa Aydın from Istanbul Technical University for their precious help on collecting and annotating the data.

6. REFERENCES

[1] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[2] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The CLEAR 2006 evaluation," in *Proc. CLEAR*. Springer-Verlag, 2006, pp. 1–44.

[3] L. Bar, S. Rochel, and N. Kiryati, "TAU-DANCE: Tel-Aviv University Multiview and Omnidirectional Video Dance Database," Vision and Image Analysis Laboratory, School of Electrical Engineering, Tel Aviv University, January 2005.

[4] A.A. Salah, R. Morros, J. Luque, C. Segura, J. Hernando, O. Ambekar, B. Schouten, and E. Pauwels, "Multimodal identification and localization of users in a smart environment," *Journal on Multimodal User Interfaces*, vol. 2, no. 2, pp. 75–91, 2008.

[5] C.H. Chen, Y. Yao, D. Page, B. Abidi, A. Koschan, and M. Abidi, "Heterogeneous fusion of omnidirectional and PTZ cameras for multiple object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1052–1063, 2008.

[6] T.E. Boult, R. Micheals, X. Gao, P. Lewis, C. Power, W. Yin, and A. Erkan, "Frame-rate omnidirectional surveillance and tracking of camouflaged and occluded targets," in *IEEE Workshop on Visual Surveillance*. 1999, pp. 48–55, IEEE.

[7] M.L. Wang and C.C. Huang, "An intelligent surveillance system based on an omnidirectional vision sensor," *IEEE Intelligent Systems*, pp. 1–6, 2006.

[8] W. Yan, C. Weber, and S. Wermter, "A hybrid probabilistic neural model for person tracking based on a ceiling-mounted camera," *Journal of Ambient Intelligence and Smart Environments*, vol. 3, no. 3, pp. 237–252, 2011.

[9] C. Vondrick, D. Ramanan, and D. Patterson, "Efficiently Scaling Up Video Annotation with Crowdsourced Marketplaces," *Proc. ECCV*, pp. 610–623, 2010.

[10] P. KaewTraKulPong and R Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Proc. 2nd European Workshop on Advanced Video Based Surveillance Systems*, 2001, vol. 25, pp. 1–5.

[11] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, May 2006.

[12] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," *Computer VisionECCV 2006*, pp. 430–443, 2006.

[13] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," *Proc. ECCV*, pp. 778–792, 2010.

[14] Koen van de Sande, Theo Gevers, and Cees Snoek, "Evaluating color descriptors for object and scene recognition.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–96, Sept. 2010.