Tracking persons using a network of RGBD cameras

George Galanakis
Institute of Computer Science, Foundation for Research and Technology Hellas
Computer Science Depart., University of Crete, Greece
ggalan@ics.forth.gr

Xenophon Zabulis
Institute of Computer Science, Foundation for Research and Technology Hellas
zabulis@ics.forth.gr

Panagiotis Koutlemanis
Institute of Computer Science, Foundation for Research and Technology Hellas
koutle@ics.forth.gr

Spiros Paparoulis
Institute of Computer Science, Foundation for Research and Technology Hellas
spriosp@ics.forth.gr

Vassilis Kouroumalis
Institute of Computer Science, Foundation for Research and Technology Hellas
vic@ics.forth.gr

ABSTRACT

A computer vision system that employs an RGBD camera network to track multiple humans is presented. The acquired views are used to volumetrically and photometrically reconstruct and track the humans robustly and in real time. Given the frequent and accurate monitoring of humans in space and time, their locations and walk-through trajectory can be robustly tracked in real-time.

Keywords

RGBD, camera network, person tracking

1. INTRODUCTION

This work presents a computer system for the detection and localization of humans in indoor environments. This system is proposed as part of an infrastructure for the emergency response to catastrophic events, such as earthquakes. Tracking of human visitors of indoor public spaces relates to the preventive actions to be taken before the event. In this respect we propose a system that employs multiple visual sensors in order to enumerate, localize and track individual persons. In this way, accurate information about the number and location of individuals exactly before the event can be provided, which is informative to rescue operations.

The unobtrusive tracking of persons calls for a platform of passive sensors that survey the environment and track humans within it. In order to cover wide areas and also to disambiguate from occlusions, multiple visual sensors are employed. Thus, real-time, robust person tracking is crucial in systems that support such interaction.

PETRA'14, May 27-30, 2014, Island of Rhodes, Greece.

ACM 978-1-4503-2746-6/14/05...$15.00

Monocular approaches to the problem [21, 22] are based on the image cues such as color and silhouette shape and employ sophisticated tracking methods to cope with scene complexity. The method in [5] utilizes a binocular camera system and combines stereo, color, shape and face detection to improve tracking performance. Though monocular approaches exhibit limitations in treating occlusions, they utilize 2D information on the appearance of persons. More recently, it was shown that humans can be tracked effectively by relying on RGB-D data [19].

Multiview human localization methods perform use a 3D representation of the imaged persons to register them to a map of the workspace. This notion of a map is reflected in several works [9, 15, 17, 7, 12, 10, 20, 23, 16], using a planar homography of the ground plane constraint to implement it. However, in order to cope with occlusions and challenging cases of tracking that occur when multiple persons occupy a space a 3D representation is required. In [11, 18, 7, 10, 20, 23, 16], a voxel grid is utilized to represent the 3D reconstruction and computation is distributed in the GPUs of four computers.

Person tracking with the help of RGB-D sensors has also been investigated recently. In [8], depth continuity of the tracked persons is assumed and combined linearly with appearance similarity. Both works [13, 1] employ three sensors in their setup. In [13] RGB-D features feed an on-line boosted system for multi-hypothesis tracking on the image plane. In [1] instead, persons are tracked from a top view, however with significant limitations to the area covered due to camera placement geometry.

Multiview tracking approaches utilize either a volumetric representation of persons or color / appearance properties as found in the image. The approach in [16] combines color and volumetric occupancy, but not seamlessly as it tracks elements by color and only disambiguates based on proximity. In this work, we combine both color and volumetric information seamlessly which exhibits tracking advantages.

2. RGBD NETWORK SETUP

A typical setup consists of a set of n RGBD cameras placed evenly and high, surrounding a volume in which persons will be localized and tracked. In order to successively track persons among different views, the data from the acquired images have to be combined, thus requiring extrinsic calibration of the sensors. We calibrate the cameras using the RGB camera of each sensor along with the common grid pattern; a large grid is placed on the floor in a place visible by all cameras and n images are acquired and calibrated using conventional grid-based calibration. The images are acquired synchronously at a given time instant, denoted as D_i and C_i, for the depth and the color image of the i_{th} sensor respectively. D_i and C_i are already rectified by the driver; actually D_i is registered to C_i. In all experiments in this paper four sensors were employed.

The above setup was designed after experimentation with possible camera placement configurations. The reason of investigation is a limitation of the network, which is met when using multiple active illumination sensors. In that case, the structured light patterns emitted from sensors may interfere and their detection by the corresponding cameras is hindered; in turn, depth estimation at the location of interference is poor. Some systems employ elaborate mechanical solution to cope with this problem. A vibrating system which reduces the interference problem [14, 4], but is practically difficult to achieve. Application of synchronized shutters, so that readings from multiple sensors are multiplexed [3], drops frame rate proportionally to the number of sensors. Due to the geometry of the proposed camera placement, pattern interference occurs on the ground plane and very little on the imaged persons. We cope with this problem using an independent estimate of the ground plane and filtering interference noise. This estimate is obtained through conventional grid based calibration along with the extrinsic calibration of the sensor network.

3. PERSON DETECTION AND TRACKING

When a set of n image pairs is acquired, persons are detected, localized and tracked in the floor plane. The detection and localization of humans is based solely on reconstructed volumes which are projected on a representation of the floor plane. In contrast, the detected persons are tracked based on both geometric and color information.

3.1 Background subtraction

Depth images contain distances from the sensor to every object in the view of the camera. In order to detect humans in this view, the background pixels must be subtracted. The foreground is computed based on depth difference from the background model, which is retained for each sensor separately. To reduce computational cost D_i images are processed in parallel, resulting to n D_i^s which contain only the depth pixels corresponding to the foreground. Fig.1 shows the calculated foreground masks for a single sensor.

3.2 3D Reconstruction

At each frame, a 3D reconstruction of the scene is obtained by concatenating the partial 3D reconstructions from the depth images D_i^s. Initially, the 3D points from the D_i^s images are transformed to triangle meshes, using their 2D neighbor connectivity to establish the vertices of the mesh. To eliminate IR interference noise at the level of the floor (see Sec. 2), triangles near the floor are filtered. Since D_i^s and C_i are aligned, the triangles are texture mapped using the registration of the depth camera to the RGB view.

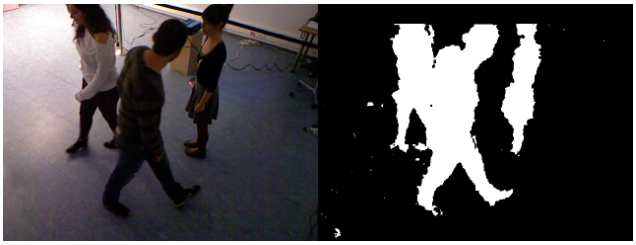

Figure 1: The input RGB channel (for illustration only) and the resulting foreground mask that was computed from the corresponding depth channel of the RGBD image.

3.3 Detection and tracking

Person representation. Following the reconstruction of the scene, the next step towards detection is to represent volumetric information in a 2D representation of the floor plane, which is similar to a map. The n views are combined in a top view of the scene rendered in GPU. This rendering is henceforth F_t, where t is the time of acquisition.

In F_t, persons are detected, represented and localized as 2D blobs. The combined image comprises the persons as if they were observed from the top, but also contains noise from the reconstruction. A binary mask F_t^b is computed for F_t and morphological operations are applied in order to eliminate such noise, but also to discriminate neighboring persons which may appear as joined. F_t^b is labeled by means of 8-connectivity and small blobs are filtered out. The detected blobs are directly localized as candidate persons on the ground plane of the scene and their centroids and textures are the measurements feeding the tracking module.

In addition, the color appearance of the person is captured. A hypothetical cylinder is assumed, centered at the location corresponding to the detected blob and enclosing the person within it. The reconstructed texture of this person is projected outwards, upon the cylinder which is then warped and 'unfolded' into a 2D image. The cylinder is approximated by an $m-gonal$ prism, and the rotation of the normal $\vec{n}$ of each facet is kept. With the use of these rotational values the blob is rendered from the m planar views, whose concatenation forms the texture associated with the blob. Examples of these, unfolded, cylindrical texture representations are shown in Fig. 2.

Color histograms represent only chromaticity of the texture and omit any information about the spatial arrangement of the color features. I.e. a person with a red top and a white shorts may not be robustly disambiguated from another who wears white top and red trousers. Moreover, if the scene has furniture it is likely that, in some cases, legs will be not fully visible. To resolve such situations, the texture is split into two parts, containing the upper and the lower body

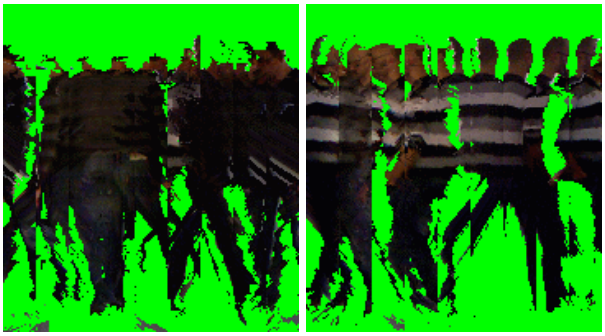

Figure 2: Two examples of unfolded cylindrical textures.

respectively. The point of intersection is statically specified and is relevant to the height of the volume from which the texture was acquired. Then two different histograms H_u^t and H_l^t are created which encode a rough information about the location of the color features. The data structure containing H_u and H_l, at time time, is referred later as A^t.

Person tracking. The blob tracker in [2] was modified to track blobs in F_t^b, rather than skin-colored blobs in color images for which it was originally developed. This tracker exhibits robustness to blob merging (in our case, due to person proximity) and transient tracking failures (often, due to lack of visibility). The tracker reinforces the hypotheses which has the stronger spatial association with the single blob. In [16], color similarity was determined in this association and in case of similar color appearance, spatial proximity was accounted. To increase robustness, the proposed tracking scheme integrates both spatial and color association in a single association metric. In addition, color association accounts for the spatial arrangement of colors on the person's surface.

The color appearance of each person is encoded in two color histograms, H_u and H_l. The samples for this histogram are collected from the texture that corresponds to the surface of this person as reconstructed in 3D. As in [16], color samples are converted to the HSV color space and finally in a 2D histogram, as only the hue and saturation components are regarded. The obtained color information is employed, to retain accurate trajectories of proximate person hypotheses. The similarity of two color histograms is quantified by a correlation metric, as in [16]. In quantifying the color cue, the obtained upper and lower histograms are compared with the corresponding histograms maintained for each tracking hypothesis. The comparison weights the similarity of upper and lower histogram pairs, for data structures A^1 and A^2:

$$d(A^1, A^2) = w_u * d(H_u^1, H_u^2) + w_l * d(H_l^1, H_l^2) \qquad (1)$$

where $k \in \{1, 2\}$, H_u^k,H_u^k are the histograms contained in A^k and w_u, w_l are adjustable weighting factors of the upper and lower body histograms. The weights are proportional to the visible area of the surface of each body segment.

In the period of a tracking loop, the tracker corresponds person hypotheses to the detected blobs. To combine the color and spatial cues the association matrix is filled as:

$$mat[i, j] = d(h_i, b_j) = d(C^{h_i}, C^{b_j}) * d(A^{h_i}, A^{b_j}) \qquad (2)$$

where i enumerates the person hypotheses h having a predicted center C^{h_i} and a texture representation A^{h_i}, while j enumerates the detected blobs b at center C^{b_j} having a texture representation A^{b_j}. Upon creation of the association matrix, the correspondences are established using an implementation of Blossom algorithm [6].

As persons are tracked through their blob representations, the associated histograms are continuously updated, in order to collect more samples of the person's surface. This update takes place if following hold:

- **Rule 1:** $d(A^{h_i}, A^{b_j})$ has a value above a threshold **OR** a person has been newly observed in the scene
- **Rule 2:** a person is not proximate to other persons which are not currently associated with a hypothesis

The first predicate is to avoid a misleading color representation, due to lack of samples , i.e. due to inaccurate reconstruction. The second predicate prevents an update of the histogram when two persons are appear as a single blob in F_t^b. If such an update was permitted, the histogram of the one could drift the appearance of the other.

At each frame that a person is tracked, as soon as the two rules are not violated, two histograms are accumulated as follows. The one that has been computed up to the current time instant and the one computed at the current frame. The accumulated histogram contains the appearance history of each person, which is sufficient in most purposes for the tracker to discriminate between persons which are temporarily lost or appear as merged and associate them correctly to blobs.

4. EXPERIMENTS

The setup for our experimentation consists of four Kinect sensors placed on floor mounts in the corners of an area sized $4 \times 4m$ and at height $\simeq 2.3m$. All sensors were connected to a single PC with an Intel Core i7, at $2.67\,GHz$ with $6\,GB$ of RAM and an NVIDIA GTX680 GPU. Each RGBD sensor has a viewing angle of $43^\circ \times 57^\circ$, a depth range of $4.5\,m$ and captures $640x480$ RGBD image pair images at $20\,Hz$. The above setup was utilized to capture a dataset including variable number of persons and means of interaction between them. Thereupon some experiments evinced the reliability of the tracker.

We performed a formative investigation focusing at known tracking limitations of using the color and spatial cues separately. In particular we focused on cases where clustered persons may share similarity on components color appearance, such as persons in Fig. 3. When two persons come proximate, they result in a single blob in the tracker's representation. In such cases due to visibility limitations, reconstruction is less complete and if person share color components, both color and spatial cues become unreliable. In this case, their combined effect was observed to perform more robustly. Fig. 3 shows tracking behavior in such a case, for the proposed method. When using the two cues independently the tracker fails to reassign the tracks correctly.

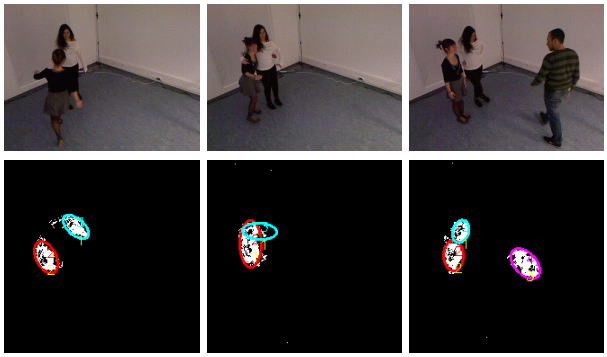

Figure 3: Two persons come close and in the middle frame appear as a single blob, but tracking retains the two hypotheses. After the blobs separate, a correct decision can be made with the use of color information; if only the predicted position is used, the IDs are incorrectly swapped.

5. CONCLUSIONS

This paper presented the use of computer vision towards the tracking of persons in an indoors environment. In the context of the emergency response for catastrophic events, future work concerns the ability of the system to operate in less constrained environments and, potentially, after some cameras have lost their initial placement location due to the event.

6. ACKNOWLEDGMENTS

This work has been supported by the FORTH-ICS RTD Programme "Ambient Intelligence and Smart Environments" and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thalis-DISFER.

7. REFERENCES

[1] E. Almazan and G. Jones. Tracking people across multiple non-overlapping rgb-d sensors. In *CVPR Workshops*, pages 831–837, June 2013.

[2] A. Argyros and M. Lourakis. Real-time tracking of multiple skin-colored objects with a possibly moving camera. In *ECCV*, pages 368–379, 2004.

[3] K. Berger, K. Ruhl, C. Brümmer, Y. Schröder, A. Scholz, and M. Magnor. Markerless motion capture using multiple color-depth sensors. In *Proceedings VMV 2011*, pages 317–324, Oct. 2011.

[4] D. A. Butler, S. Izadi, O. Hilliges, D. Molyneaux, S. Hodges, and D. Kim. Shake'n'sense: Reducing interference for overlapping structured light depth cameras. In *SIGCHI*, pages 1933–1936, 2012.

[5] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. In *CVPR Workshops*, pages 601–609, 1998.

[6] J. Edmonds. Paths, trees, and flowers. *Canadian Journal of mathematics*, 17(3):449–467, 1965.

[7] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *PAMI*, 30(2):267–282, 2008.

[8] J. Han, E. Pauwels, P. de Zeeuw, and P. de With. Employing a rgb-d sensor for real-time tracking of humans across multiple re-entries in a smart environment. *Consumer Electronics, IEEE Transactions on*, 58(2):255–263, May 2012.

[9] S. Khan and M. Shah. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *ECCV*, pages 133–146, 2006.

[10] K. Kortbek and K. Gronbaek. Interactive spatial multimedia for communication of art in the physical museum space. In *ACM Multimedia*, pages 609–618, 2008.

[11] A. Ladikos, S. Benhimane, and N. Navab. Efficient visual hull computation for real-time 3D reconstruction using CUDA. In *CVPR Workshops*, pages 1–8, 2008.

[12] M. Liem and D. Gavrila. Multi-person tracking with overlapping cameras in complex, dynamic environments. In *BMVC*, 2009.

[13] M. Luber, L. Spinello, and K. O. Arras. People tracking in rgb-d data with on-line boosted target models. In *IROS*, pages 3844–3849. IEEE, 2011.

[14] A. Maimone and H. Fuchs. Reducing interference between multiple structured light depth sensors using motion. In *VRW*, pages 51–54, 2012.

[15] A. Mittal and L. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. In *IJCV*, pages 189–203, 2003.

[16] P. Padeleris, X. Zabulis, and A. A. Argyros. Multicamera tracking of multiple humans based on colored visual hulls. In *ETFA*, pages 1–8, 2013.

[17] D. Reddy et al. Compressed sensing for multi-view tracking and 3-D voxel reconstruction. In *ICIP*, pages 221–224, 2008.

[18] A. Schick and R. Stiefelhagen. Real-time GPU-based voxel carving with systematic occlusion handling. In *DAGM Symp.*, pages 372–81, 2009.

[19] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

[20] F. Sparacino. Scenographies of the past and museums of the future: from the wunderkammer to body-driven interactive narrative spaces. In *ACM Multimedia*, pages 72–79, 2004.

[21] S. Tran, Z. Lin, D. Harwood, and L. Davis. UMD VDT, an integration of detection and tracking methods for multiple human tracking. In *Multimodal Technologies for Perception of Humans*, pages 179–190. Springer, 2008.

[22] B. Wu, V. Singh, C. Kuo, L. Zhang, S. Lee, and R. Nevatia. CLEAR'07 evaluation of usc human tracking system for surveillance videos. In *Multimodal Technologies for Perception of Humans*, pages 191–196. Springer, 2008.

[23] X. Zabulis, D. Grammenos, T. Sarmis, K. Tzevanidis, P. Padeleris, P. Koutlemanis, and A. A. Argyros. Multicamera human detection and tracking supporting natural interaction with large-scale displays. *Mach. Vis. Appl.*, 24(2):319–336, 2013.