A multicamera vision system supporting the development of wide-area exertainment applications

Xenophon Zabulis, Thomas Sarmis, Dimitris Grammenos, Antonis A. Argyros
Institute of Computer Science – FORTH
Herakleion, Crete, Greece
{zabulis,sarmis,gramenos,argyros}@ics.forth.gr

Abstract

In this paper, the application of computer vision techniques to the localization of multiple persons in a relatively wide gaming terrain is presented. Multiple views are employed both for terrain coverage, but most importantly, for treatment of occlusions. Through the appropriate selection of lightweight operations and acceleration strategies, an adequate frame rate is achieved despite the large volume of input data. The resulting system is employed in the development of multiplayer entertainment applications, which are demonstrated and evaluated.

1. Introduction

Advances in visual interpretation of human motion, have given rise to interaction models beyond the traditional paradigm, in which a video game is played by a user sitting in front of a computer display. Off-the-shelf gaming and entertainment systems already exist, in which user input is provided through hand and body motion. Such systems constitute examples of an emerging trend in digital entertainment that has been dubbed as "exertainment", because it capitalizes on the use of video games to enhance and stimulate physical exercise [13].

Currently, off-shelf vision-based applications feature single cameras operating in restricted spatial ranges. Due to lack of 3D information, monocular vision approaches provide limited interaction capabilities. Furthermore, there is no or very little support for multi-user interaction, mainly due to the difficulty of interpreting images where users occlude each other. To date, multi-user involvement in a wide interaction terrain is enabled through approaches where 3D information is obtained by inertial (e.g. Nintendo's *Wii*), or tactile (e.g. Konami's *Dance Dance Revolution*) sensors, therefore requiring the user to carry or wear an additional device.

This work employs state-of-the-art computer vision techniques towards the development of wide-area, multiplayer, and non-invasive exertainment applications. Multiple views are employed to robustly provide granular 3D information about multiple persons in wide areas, and at an adequate frame rate. The efficacy of the proposed approach is evaluated, through a handful of pilot, or "mini-game" demonstrators.

The remainder of this paper is organized as follows. Related work is reviewed in Sec. 2. The setup of the employed hardware infrastructure is presented in Sec. 3 and the computer vision approach in Sec. 4. In Sec. 5, the proposed approach is used to support the development of several mini-game applications, which are presented and discussed. Finally, in Sec. 6, this work is summarized and directions for future work are provided.

2. Related Work

To date, there exist two main paradigms for the use of computer vision techniques to video game applications. The first replicates the user's appearance and embeds it in the world of the game. The second mimics and, sometimes, interprets user motion to control game entities (i.e. player avatars).

In one of the earliest approaches, the *KidsRoom* [1], an interactive children's bedroom guides the audience through an adventure story. *KidsRoom* employed computer vision for object tracking, motion detection and simple action recognition. Game feedback was provided through back-projected video displays, speakers and theatrical illumination.

More recently, generic game consoles have featured a collection of games that utilize a camera as an interaction enabler. In games for the Microsoft's *Xbox 360*, an image of the user's face is superimposed on the face of an in-game avatar. In a few other titles, coarse player motion is utilized as a game controller. More elaborate in this aspect is Sony's *EyeToy* extension for the *Playstation* console, which detects the user's silhouette from a frontal view and replicates it in the virtual gaming terrain. Through this representation, the user can hit virtual targets or perform physical exercises in the world of the game, by performing the equivalent actions in the physical world. By measuring optical flow, motion is exaggerated in [4], in a martial arts game. In addition, large-scale and surround displays are employed to increase the sense of immersion into the world of the game.

In the above, video input is limited to coarse 2D body motions or gestures, there is no 3D or depth information, the player is required to stand at approximately the same distance from the camera in a restricted spatial region, while multiplayer interaction (when available) requires that players do not occlude each other in the image. The systems in [1, 5] utilize one or more a vantage viewpoints from the ceiling, to map motion of multiple players on the 2D map of the game. Recognized coarse body gestures trigger gaming actions (i.e. jump, fire). The game terrain is projected on a wall by two adjacent video projectors.

The computational process that supports the majority of these games is background subtraction. This process is employed to segment the foreground objects (persons) from the image. Based on this segmentation, the user's silhouette and location are estimated. In multiview systems, background subtraction has been utilized to obtain 3D information about the imaged persons. By projectively combining the obtained silhouettes, a 3D reconstruction of the imaged person(s) is obtained through

the visual hull [7]. Skeletonization of this hull is employed in [9, 11] to recognize body posture, while in [10], the body parts are piecewise tracked through surface registration.

3. Hardware setup

The system is installed in a 5,05×4,80 m^2 room with a large display on each wall (see Fig. 3a). These displays are two 32" HDTV screens, a 60" semi-transparent back-projection screen, and a projection wall covered by two adjacent projectors. Each display is equipped with stereo speakers and an additional 8-speaker 3D surround audio system is installed on the ceiling.

The utilized vision system is comprised from eight FireWire cameras ($66^o \times 51^o$) that are mounted at the corners and at the in-between mid-wall points of the room viewing it peripherally in steps of 45^o. At the same height, two additional cameras are mounted on the ceiling, with their optical axes normal to the floor. Cameras are synchronized by a dedicated FireWire bus.

Cameras provide views of the room, henceforth indexed by **i**. Cameras are intrinsically calibrated and extrinsically registered utilizing the toolbox in [2]. The estimates of extrinsic parameters are then refined using the bundle adjustment method in [6]. The camera projection matrices are denoted by $\mathbf{P_i}$.

The system is supported by a custom software platform that enables synchronous image acquisition and facilitates their on-line distribution across multiple computers and processes.

4. Computer vision processes

At each frame, the acquired images are rectified for lens distortion and then background subtracted [14], to yield a set of corresponding binary images. These images are typically noisy, due to errors in the background subtraction process. Such errors typically occur as small holes or intrusions in the silhouette of the person; in addition, small background regions are spuriously detected as foreground. To compensate for these errors, a sequence of morphological operations (open, close, dilate, erode) is applied to each binary image. The result is binary images $\mathbf{B_i}$, in which a pixel has a value of 1 if it belongs to the foreground and 0 otherwise.

Images $\mathbf{B_i}$ are combined into a volumetric representation, which is comprised of the visual hull(s) of the imaged person(s). This representation utilizes a volumetric occupancy grid **V**, in which a voxel is eventually assigned with the value of 1, if it is considered to be occupied by a person and 0 otherwise. In all experiments in this paper, **V** was tessellated in voxels of $1cm^3$ and was aligned with the room. **V** covers totally the floor of the room, and reaches a height of $2.5m$.

Values are assigned to voxels in **V** based on the observation that if a voxel is occupied by a foreground object (person), its projection will occur in foreground regions in *all* images $\mathbf{B_i}$ [7]. Let **i'** index the views where voxel **v** occurs within their visual field. The value of **V** at **v** is computed by a logical *AND* operation across the values of $\mathbf{B_i}$ at which **v** projects:

$$\mathbf{V}(\mathbf{v}) = AND_{\mathbf{i'}} (\mathbf{B_{i'}} (\mathbf{P_{i'}}, \mathbf{v}))$$

The above approach is robust against false-positives in the background subtraction process (i.e. cases where a background pixel has been classified as foreground). The reason is that, due to the multiplicative nature of the *AND* operation, such an error must co-occur at corresponding image locations of *all* remaining views in order to give rise to an occupied voxel, which is highly unlikely. On the other hand, a false-negative error in just one view will spuriously reduce the volume of the visual hull. Therefore, background subtraction is tuned to favor false positive detections of foreground pixels versus false negative ones. The combination of this strategy with the morphologic processing of the background subtraction results provides adequately realistic 3D models.

Finally, the visual hull of each individual person is extracted as a 3D blob in **V**, by a connected component labeling (CCL) process. Attributes of the visual hull of each person are computed in each frame and updated on-line, to serve as input to the exertainment application. It is worth noting that since the proposed approach is based on multiple views it can cope with occlusions, as long as, there are enough views to disambiguate between persons.

A basic attribute of each hull is its 3D bounding box, which provides a practical approximation of the volumetric region occupied by a person. This approximation is utilized to track persons, by corresponding overlapping bounding boxes across consecutive frames. Fig. 1 illustrates the above operations for a particular set of four views. Note that, although three persons have been correctly identified in the volumetric representation (bottom left), persons are occluded in all four 2D camera views.

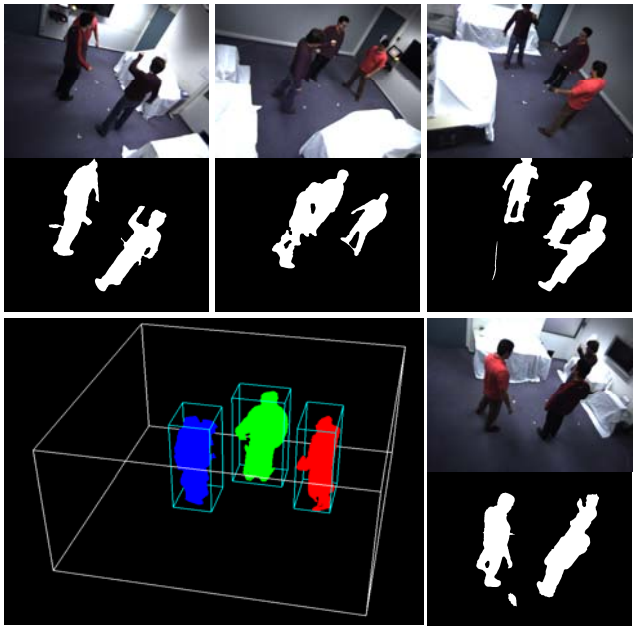

Figure 1. Visual hull extraction and person localization (bottom left). The original images and the morphologically processed foreground segmentation images $\mathbf{B_i}$ are also shown.

The centroid of a person's visual hull provides the 3D location of the corresponding person in the room. To provide with a robust estimation of the persons' locations, a Kalman filter [8] is assigned to each of them. The projection of the estimated 3D centroid on the room's floor provides the coordinates of a person in the 2D game terrain.

The footprints of the person in the floor can also be

computed as the intersection of a person's hull with the floor plane. In particular, the values of voxels in **V** for which $\mathbf{z} = 0$ are utilized to form an image. In this image, footprints are extracted as blobs by a CCL process. This way, a footprint is obtained only when the foot of a person is in contact with the floor. Finally, footprints are associated with persons through the particular 3D bounding box within which they occur. In Fig. 2, footprint extraction is demonstrated for the presice dancing figures of a person, in a reference dataset[1].

Figure 2. Footprint extraction. The obtained footprints over 20 frames are shown superimposed in an original image (left) and drawn on a virtual canvas (right), as would appear from a top view.

To be able to perform the necessary computations at an acceptable frame rate, the above operations are parallelized and pipelined. Each video stream originating from a camera is assigned to a separate CPU that performs background subtraction and morphologic image filtering operations. The output images $\mathbf{B_i}$ are run-length encoded to be efficiently transmitted to another CPU for the computation of the visual hull. Once this transmission is complete the original CPU initiates the background subtraction processing of the next frame.

The computation of the visual hull is also performed in parallel, by partitioning voxel space **V** and assigning of each partition to a different CPU. Within each partition, the computation is accelerated by performing the *AND* operation in a "view-first" order. Specifically, the operation initially evaluates all voxels in **V** for the first two views and stores this intermediate result in a voxel space **V'**. For each remaining view, the *AND* operation is performed only for the voxels for which **V'** is 1 and stores the result in **V'**, overwriting it. The last **V'** is copied back to **V**.

In the implementation, 8 CPU cores are employed in total, equally distributed in two computers. A dedicated local area network link of 1GB bandwidth is reserved for their communication. A frame rate of 10 *Hz* has been achieved when employing 4 views in a resolution of 960×1280 pixels and a **V** of 5×5×2.5 m^2 tessellated in voxels of *1* cm^3. Increasing the number of employed views to 8 while using the same number of CPU cores, reduces the frame rate to 8 *Hz*. Note that given more CPU cores the level of parallelization and, thereby, the frame rate can be increased.

In practice, it was observed that four cameras provide adequate coverage for 3 persons it the above area. More than four cameras are employed when players are more than 3, or when increased localization accuracy is required (e.g. in the footprint scenario). The error in localization is in the order of 5 *cm* for 4 cameras. In the footprints application, 8 cameras allow for finer user-control when drawing on the virtual canvas.

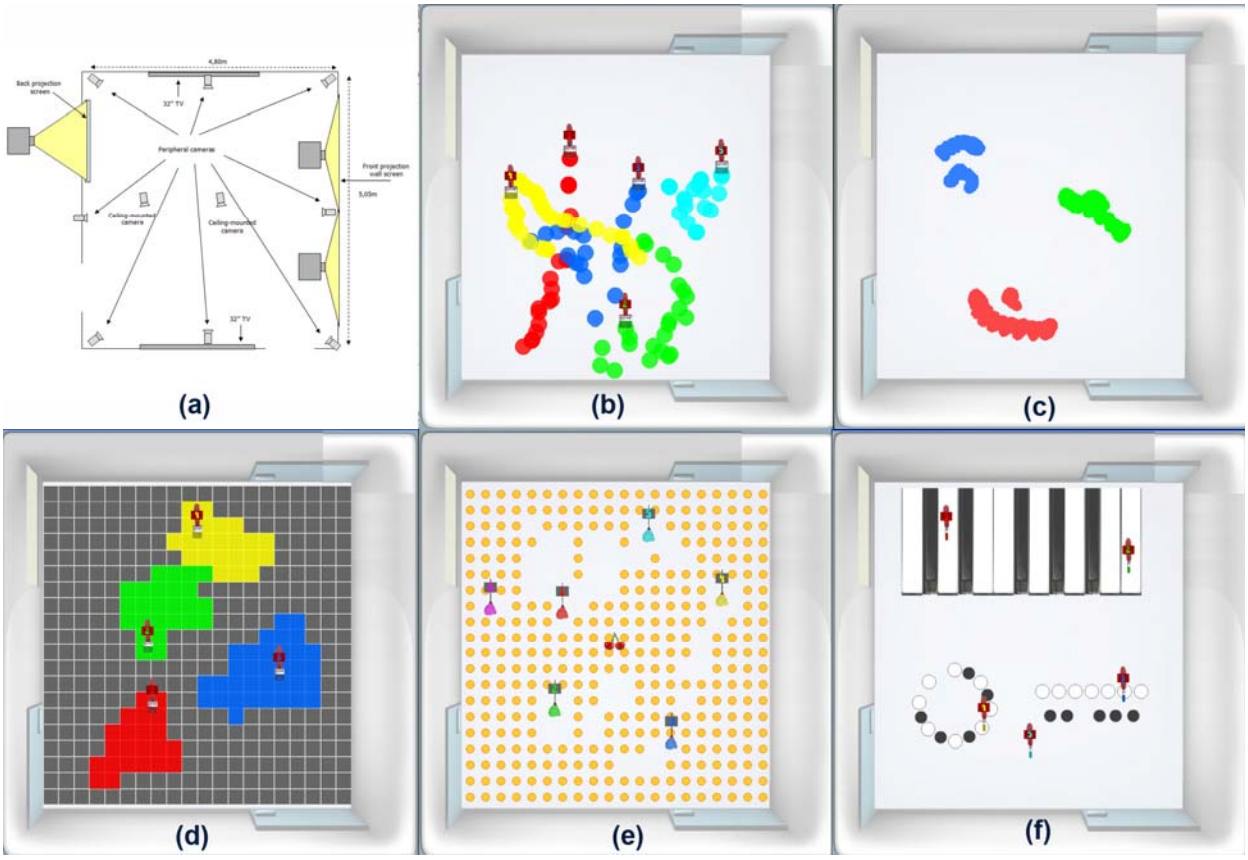

Figure 3. (a) Room layout; (b) – (f) The user interfaces of the mini-games.

[1] Downloaded from https://charibdis.inrialpes.fr

5. Experiments with the mini-games

To test the effectiveness of the vision infrastructure in supporting exertainment scenaria, 5 mini-games have been developed. The games follow a client-server architecture. The server communicates with the vision subsystem and then provides on-line location information to the clients which render the game's visual interface. An unlimited number of clients can connect to the server. In our test case, one client was presented on each available display, so that players could always have a view of the game irrespectively of their orientation.

In the games, the user interface background presents a top-down view of the room ("virtual floor"). Some landmarks of the room are displayed to facilitate visual comprehension of this map. Furthermore, the interface can be rotated, so that it can be accurately aligned relatively to the room, depending on the position of the screen in which rendering is performed. Players are represented as distinctively colored brushes (or, in one case, mops). Whenever new persons enter the room, they are automatically added as new players. The developed games are the following:

Body painting (see Fig. 3b)**:** As players move around the terrain, they paint on the virtual floor. Using a mobile phone as a remote control, users can change properties of the brush associated to them. The created paintings can be printed or saved as JPEG images.

Footprints (see Fig. 3c)**:** As players move around, the contact area of their feet with the floor is used to paint on the virtual floor.

Paint the floor (see Fig. 3d): The floor is divided in a 20x20 grid of grey tiles. As players move around, they paint with their color the tiles that they are stepping on. The winner is the player that will manage to paint first a number of tiles that is equal to the total number of tiles divided by the number of players.

Floorman (see Fig. 3e): Players, represented by mops, have to clear the dots off the virtual floor to gain score points. A bonus fruit (e.g., a cherry) randomly appears on the virtual floor's center. The player who reaches it becomes a hunter for a few seconds and can steal points from other players by colliding with them.

Walkman (see Fig. 3f): Alternating black and white areas representing piano keys are laid on the floor. Players can create music by stepping on them.

The mini-games were evaluated by and showcased at a large number of subjects (players and audiences). Both subjects and audiences exhibited great diversity in age, gender, cultural and professional background and were all naïve to the experimental hypotheses.

The overall impression was that the games are considered as exciting and engaging, by a quite diverse audience. Through observations and discussions, it was concluded that what players enjoyed most about the games was that: (a) their game play was easy and intuitive, (b) there were no cables or cumbersome controllers and (c) new players could dynamically join the game. Furthermore, almost everyone commented positively on the fact that the games encouraged players to engage physical exercise. On the other hand, lack of visual feedback on the floor was identified, since occasionally players could not immediately identify their position on the virtual floor. This effect was more intensely pronounced when a player's view towards a display was occluded by another person.

6. Discussion

This paper presented the development of a multiview system that granularly localizes multiple persons and extracts geometric attributes of these persons, in wide areas. This system plays the role of a computer vision infrastructure for the development of exertainment applications which combine video gaming with physical exercise. To demonstrate the effectiveness of the approach, a set of gaming demonstrators utilizing this platform has been developed and evaluated.

The contribution of this work is the employment of state-of-the-art multi-view computer vision techniques in the field of gaming applications, in order to support multiplayer interaction in wide gaming areas (terrains). In addition, the use of footprints in entertainment applications is, to the best of our knowledge, novel.

Ongoing research and development efforts are focused on extracting enhanced descriptions of player motion and interaction to support richer interaction of the players with the application. In this direction, efforts are focused on increasing the detail of extracted representations (e.g. detecting head and hand locations), and including dynamic attributes such as velocity and significant changes in player body posture (e.g. sitting, ducking, falling).

Acknowledgements: This work has been partially supported by the FORTH-ICS internal RTD programme 'AmI:Ambient Intelligence Environments'.

References

[1] A. Bobick et al, The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment. Presence: Teleoperators and Virtual Environments 8(4):369-393, 1999.

[2] J. Bouguet, Camera Calibration Toolbox for MATLAB, http://www.vision.caltech.edu/bouguetj/calib_doc/

[3] D. Comaniciu and P. Meer. Mean Shift: A robust approach toward feature space analysis. PAMI, 24:603–619, 2002.

[4] P. Hamalainen, T. Ilmonen, J. Hoysniemi, M. Lindholm, and A. Nykanen. Martial arts in artificial reality. In CHI, pages 781-790, 2005.

[5] S. Laakso and M. Laakso. 2006. Design of a body-driven multiplayer game system, Computers in Entertaintament 4(4):7-4C, 2006.

[6] M. Lourakis and A. Argyros, SBA: A Software Package for Generic Sparse Bundle Adjustment, ACM Transactions on Mathematical Software, 36(1), 2009.

[7] T. Matsuyama, X. Wu, T. Takai, and S. Nobuhara, Realtime 3D shape reconstruction, dynamic 3D mesh deformation, and high fidelity visualization for 3D video. CVIU, 96(3):1077–3142, 2004.

[8] P. Maybeck, Stochastic models, estimation, and control, Mathematics in Science and Engineering, vol 141, 1979.

[9] C. Menier, E. Boyer, and B. Raffin. 3D skeleton-based body pose recovery. In 3DPVT, pages 389–396, 2006.

[10] L. Mundermann, S. Corazza, and T. Andriacchi. Markerless human motion capture through visual hull and articulated ICP, In NIPS, 2006.

[11] C. Theobalt, M. Magnor, P. Schuler, and H. Seidel. Combining 2D feature tracking and volume reconstruction for online video-based human motion capture. In Pacific Graphics, page 96, 2002.

[12] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In CVPR, pages 511–518, 2001.

[13] C. Waine. TV/video games and child obesity. Trust for America's Health Annual Report, September 2007.

[14] Z. Zivkovic. Improved adaptive Gaussian mixture model for background subtraction. In ICPR, pages 28–31, 2004