Video-Based Multi-person Human Motion Capturing

Nico van der Aa[1], Lucas Noldus[1], Remco Veltkamp[2]

[1]*Noldus Information Technology, Wageningen, The Netherlands. n.vanderaa@noldus.nl; l.noldus.noldus.nl*

[2]*Utrecht University, Utrecht, The Netherlands. R.C.Veltkamp@uu.nl*

Abstract

Observing people using standard cameras provides a non-intrusive way to capture human motion for further study of human behavior in a scene. The system presented uses multiple cameras to detect subjects, track them over time and gives 3D insights in how a subject moves around. These cues are important for the interpretation of human behavior in response to its environment, to interaction with other people or with interactive systems.

Introduction

To capture people's behavior in specific studies is complex. To illustrate, in the Restaurant of the Future 1, researchers want to study people's behavior with respect to food selection and food consumption. Knowing how people choose their food, how other people influence this choice, and if they like what they eat are some of the research questions. The Restaurant of the Future is a field lab equipped with 23 pan-zoom-tilt cameras to ensure non-intrusive observations for food selection and consumer behavior studies. Most studies use The Observer® XT 2: an event logging software for the collection, analysis, and presentation of observational data. Because of the diversity in study setups, full automation is nearly impossible. Therefore, a software module is created based on computer vision algorithms called the Video Analysis and Recognition Toolbox (VidART) to help researchers to focus on the important parts. The tedious manual annotations from video streams are partly reduced by using vision-based algorithms to detect and track people from video streams. To illustrate this, imagine you want to know how often your product is chosen. By tracking the subjects' position automatically and knowing your product's position, the system can select only those frames for further analysis, where people actually approach your product.

Video Analysis and Recognition Toolbox (VidART)

Video analysis in the Restaurant of the Future is challenging due to complex lighting conditions (big windows, different illumination possibilities, large shadows, etc.), the unconstrained subject's appearance (wide clothes, reflective jewelry, hair fashion, etc.) and the presence of static objects in the scene. VidART consists of four modules: (1) background subtraction and shadow detection; (2) contour tracking; (3) voxel reconstruction; and (4) people position tracking.

Background subtraction techniques, like the Mixture of Gaussians 3, are used to detect people or objects in a camera view. The key idea is that any object present in the current frame, but absent in the background model, is identified as foreground. This background model is computed from one or more previous frames captured by the same camera. A subject's silhouette provides essential information about its shape, and other features are easily derived from it like its appearance. It forms the basis for further analysis like tracking and pose estimation. Although background subtraction gives these silhouettes, it requires static cameras and it is sensitive to illumination changes and shadows. Shadows can be detected by assuming that shadows do not alter the texture of the object or the color chromaticity. However, shadows will also be present inside the silhouettes, since the light source's position is in general not the same as that of the camera. Removing such shadows will cause undesirable holes in the silhouette, leading to severe problems in the voxel reconstruction.

An alternative to image segmentation using background subtraction can be found in segmentation based on evolving a so-called level set function. Two appearance models are used: one for the background appearance and one for the foreground appearance. These models, each consisting of a histogram computed from an initial

segmentation, are used to drive the segmentation towards the boundary of the desired object. It has the same goal as background subtraction, but it removes the disadvantages. Unfortunately, it is very slow in general. Since we are using video streams, the segmentation result from the previous frame can be used to speed up the process. Our level set tracker based on 4 goes even one step further. It distinguishes rigid body motion from deformations. The rigid body tracking assumes that the shape does not change and locates the position of this rigid body in the next frame. This is represented by a warp of a bounding box, keeping the segmentation inside fixed. The deformation phase finds an optimal segmentation in the bounding box by using the level set segmentation. Although computationally expensive, it is reduced to a small region-of-interest, making the overall method fast. Because of the distinction between rigid body movement and shape deformations some additional correction might be necessary, since deformations can also include rigid body movements. Therefore, an additional drift correction is computed to keep the segmentation in the center of the bounding box. Figure 1 shows the steps of the level set based tracker applied to track a person's head.

Once the silhouettes are obtained for each camera view, we can project the 2D foreground silhouettes to the 3D world. The technique used is voxel reconstruction 5, which keeps a registration of which pixel in each camera view belongs to which voxel (the 3D variant of a pixel) in the 3D world. These pixel-to-voxel correspondences can only be obtained if the cameras are calibrated. In other words, we have to know how a 3D point is captured by the camera and projected to the 2D image. A common way to have the cameras calibrated is to use a calibration object, like a checkerboard, of which the size and dimensions are known 6. Voxel reconstruction labels a voxel as "visible", if all corresponding pixels in the camera views are indicating a foreground object. As a result, a 3D reconstruction of the person's or object's silhouette is available. In Figure 2 an example is given.

To track a person, the silhouettes from either background subtraction or level set based tracking are used to estimate the position of each person in the scene. Initially, a histogram is created for each person to capture his/her appearance. In 7 the concept of vertical reference lines is explained, which are the 2D correspondences of selected 3D lines perpendicular to the ground floor projected on the image planes. For each vertical reference line, the pixel count is registered for each person. For each of the pixels on this vertical reference line that are classified as being foreground pixels, the probability is computed that this pixel belongs to person *k*. The pixel counts for the person with the highest probability. The vertical reference line with the highest pixel count for person *k* is taken as his/her position in the 2D view. Once these lines are found in all cameras views, they are projected back to 3D. To handle occlusions, like when a person is in front of another person in a certain view, the concept of Best Visibility View is used 7. This concept uses the principle that the back-projection of the lines only is done using those cameras that see the person best. Figure 3 shows an example where each person is best visible in two cameras. This makes the tracking of people more robust.

Figure 1. Principle of level set based contour tracking: the bounding box (green) including the initial segmentation (yellow) are given in time step *t*. In the next time step *t*+1, the rigid body tracking, the re-segmentation and the drift correction step are given.

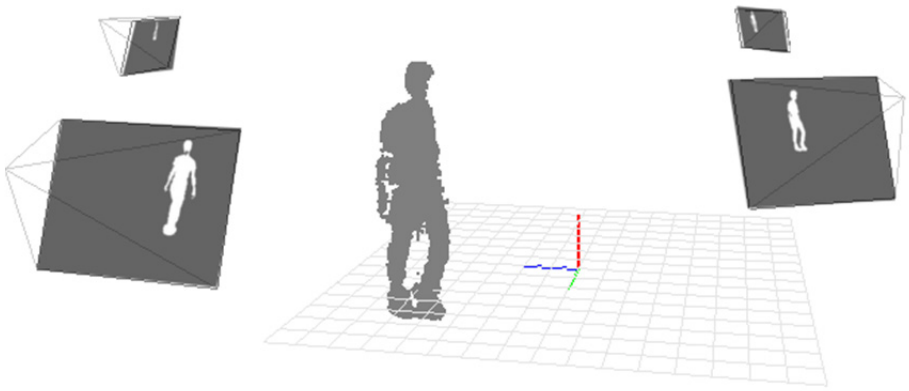

Figure 2. Example of how foreground detection of multiple cameras are combined to obtain a voxel reconstruction of the foreground objects.

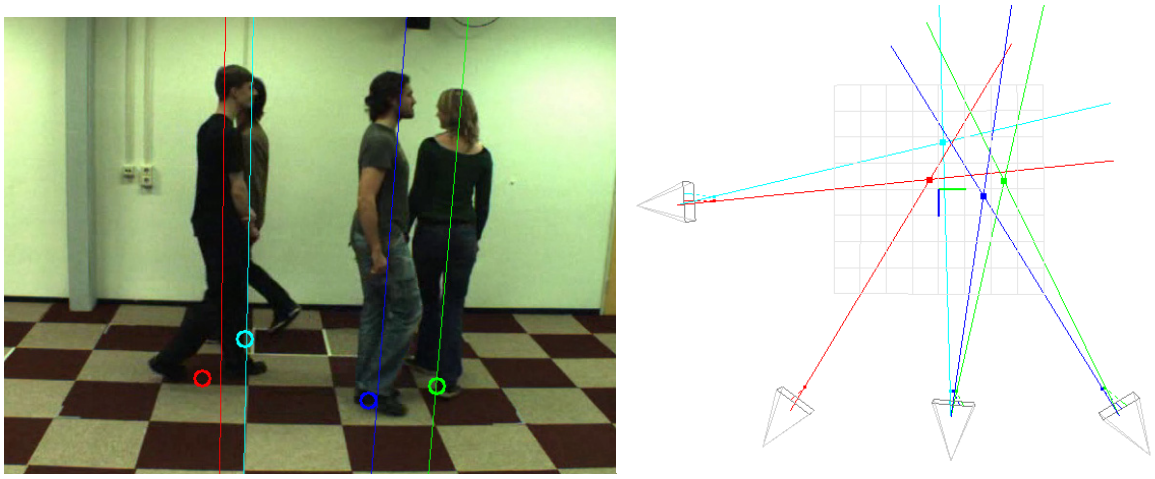

Figure 3. Principle of people tracking with vertical reference lines (*left*) and a top view of the ground floor position using the best visibility view (*right*).

Validation

The software modules have to be tested on accuracy and speed. Since the Restaurant of the Future is a challenging environment, we created the Utrecht Multi-Person Motion (UMPM) benchmark 8 to provide a benchmark for articulated human motion capturing for multi-person motion and interaction in a more controlled environment. UMPM exists of scenarios including human-human and human-object interaction, where each scenario is recorded with up to 4 subjects in the scene. For each scenario, UMPM includes synchronized video sequences for the four cameras used with 644 x 484 resolution at 50 fps. The scenarios include subjects that (1) walk, jog and run in an arbitrary way among each other, (2) walk along a circle or triangle of a predetermined size, (3) walk around while one of them sits or hangs on a chair, (4) sit, lie, hang or stand on a table or walk around it, and (5) grab objects from a table. These scenarios include individual actions, but the number of subjects moving around in the restricted area cause inter-person occlusions. We also include two scenarios with interaction between the subjects: (6) a conversation with natural gestures and (7) the subjects throw or pass a ball to each other while walking around. Since the UMPM benchmark is created to capture human motion, 3D ground truth information is provided, knowing marker positions captured by a motion capture system. The UMPM benchmark including the videos, background images, calibration data and ground truth 3D information, is available for research purposes on www.projects.science.uu.nl/umpm/. In the figures shown in this paper the UMPM benchmark is used to illustrate the principles of the toolbox.

Conclusions and future work

The VidART software ensures people tracking in 3D from video streams, which will be an important tool for automatic detection of people's behavior. Although it is currently tested on the UMPM benchmark, which is a controlled indoor environment, the toolbox will be extended to be applicable in more advanced sceneries like the Restaurant of the Future in the near future.

Acknowledgement

This research has been supported by the GATE project, funded by the Netherlands Organization for Scientific Research (NWO).

References

1. Wageningen UR (2011). Restaurant of the Future, <http://www.restaurantvandetoekomst.wur.nl/UK/>, Accessed 27 March 2012.

2. The Observer XT software, < http://www.noldus.com/observer>, Accessed 27 March 2012.

3. Stauffer, C., Grimson, W.E.L. (1999). Adaptive background mixture models for real-time tracking, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 1999*, 246-252.

4. Bibby, C. Reid, I. (2008). Robust Real-Time Visual Tracking using Pixel-Wise Posteriors, *Proceedings of the 10th European Conference on Computer Vision, Marseille, France, October 2008*, Part II, 831-844.

5. Kehl, R., Bray, M., Gool, L. van (2005). Full body tracking from multiple views using stochastic sampling, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 2005*, 129-136.

6. Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(11): 1330-1334.

7. Luo, X., Tan, R.T., Veltkamp, R.C. (2011). Multi-person Tracking Based on Vertical Reference Lines and Dynamic Visibility Analysis. *IEEE International Conference on Image Processing (ICIP 2011)*.

8. Aa, N.P. van der, Luo, X., Giezeman, G.J., Tan, R.T., Veltkamp, R.C. (2011), Utrecht Multi-Person Motion (UMPM) benchmark: a multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction, *Proceedings of the Workshop on Human Interaction in Computer Vision (HICV), in conjunction with ICCV 2011*.