Monitoring Human Activities: Flexible Calibration of a Wide Area System of Synchronized Cameras

Stéphane Drouin, Régis Poulin, Patrick Hébert and Marc Parizeau
Computer Vision and Systems Laboratory
Department of Electrical and Computer Engineering
Laval University, Sainte-Foy, QC, Canada, G1K 7P4
{sdrouin, poulin06, hebert, parizeau}@gel.ulaval.ca

Abstract

This paper presents an efficient procedure for calibrating a wide area system of synchronized cameras with respect to a global coordinate system. The flexibility of this procedure lies in its use of a hand-held calibration target that is automatically detected in real-time by the system, and also in its capability to provide a feedback on the working volume coverage through a sampling pyramid. In a first stage, the intrinsic parameters are computed separately for each camera, and then the relative positions of the cameras are evaluated through a bundle adjustment. The 3D reconstruction uncertainty of the camera system is assessed locally within the working volume and is verified to be acceptable for monitoring human activities.

1 Introduction

In order to conduct experiments on automatic description of human activities from a set of multiple cameras, one needs an appropriate experimental system. The system should be transportable and easily reconfigurable to adapt to the environment where the experiment is held. Typically, an experiment will be performed in an area the size of a room. Since the idea is to capture dynamic 3D features of the scene in real-time, it is necessary to synchronize the video streams [5, 7] and calibrate the camera parameters [3, 9]. While the calibration mathematical model is chosen from precision requirements (1cm in a room-sized volume), the calibration procedure must be rapid and kept simple. Moreover, depending on the configuration of the cameras, the local uncertainty of the 3D measurements may vary significantly within the observed volume (working volume). It is thus important to provide an indication of the expected local uncertainty after calibration. This uncertainty is then available to the application when 3D reconstruction is performed.

To calibrate a single camera, Zhang [9] proposed a method using a planar pattern that is easily moved in the camera's field of view; the intrinsic parameters as well as a set of extrinsic parameters for each position of the cameras are estimated. Heikkilä [3] improved the accuracy of the calibration using a more elaborate model for distortion. Nevertheless, these approaches do not integrate a systematic method to calibrate a set of cameras with large fields of view. For instance, there is no indication on sampling the working volume (WV) using the calibration pattern. Surprisingly, very few calibration approaches for wide area systems have been proposed [1, 8]. Although these approaches differ in their implementation complexity, they generally decouple the estimation of intrinsic and extrinsic camera parameters. Intrinsic parameters are first estimated for each camera independently. Secondly, the relative position of the cameras is calibrated by using a fixed or moving calibration target in the WV. For fixed environments, it is possible to integrate landmarks and survey them (e.g. using a theodolite) before calibration [8]. However, this method cannot easily accommodate a transportable system. Interestingly in [1], a single LED point is waved in front of the camera network to calibrate the extrinsic parameters. In this case, the LED need not be seen simultaneously by all cameras. Nevertheless, since the system is not synchronized, assumptions must be made to decouple spatial and temporal errors. While segmentation of the LED is simple for indoor environments, this approach cannot function outdoors. Moreover, the small size of the LED's image does not allow the position estimate to be precise. On the other hand, a calibration pattern formed of circular patches can easily be segmented within a tenth of pixel precision. In order to do so, the diameter of a circular patch image should be approximately 20 to 30 pixels.

This paper presents a flexible calibration procedure for a wide area system of synchronized cameras. The presented procedure allows for fast calibration using hand-held planar targets that can be moved freely throughout the WV. A method for sampling the WV with the cali-

bration pattern both for intrinsic and extrinsic parameters calibration is presented. The configuration of the cameras can thus be changed and calibrated within minutes. Also, since the calibration targets provide many points, it is possible to locally assess the quality of the reconstruction within the WV.

The paper is organized as follows: Section 2 presents the calibration procedure, Section 3 describes the procedure to assess the quality of the calibration and experimental results are given in Section 4.

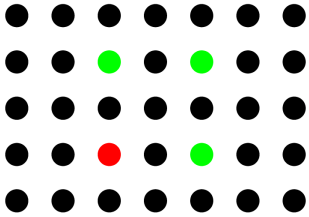


Figure 1: Calibration target (30 or 60 cm long).

2 Calibration procedure

The calibration procedure begins with a separate estimation of the intrinsic camera parameters. For this purpose, the geometric camera calibration algorithm described by Heikkilä [3] is used. In the second step, global extrinsic parameters are estimated from homography [9] and a bundle adjustment is performed to optimize all of the parameters.

We use the **pinhole camera** model for 3D reconstruction. The pinhole models the projection of a 3D point $\mathbf{w} = (x, y, z)$ into an image point $\mathbf{m} = (u, v)$ by a projection matrix $\mathbf{P} = \mathbf{A}[\mathbf{R}|\mathbf{t}]$ that represents the intrinsic and extrinsic parameters of the camera. The intrinsic parameters $\mathbf{A}$ relate the camera coordinates (millimeters) to the image coordinates (pixels). The extrinsic parameters relate the global coordinate system to the camera reference frame. They are represented by a rigid transformation (rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$). Nonlinear distortions are also added into the model in order to deal with short focal length lenses that are typically used in the applications we are interested in. The model includes both radial and tangential distortion coefficients [3].

Calibration targets These targets are composed of filled circles arranged on a 5×7 rectangular grid (see Figure 1). Similar patterns of different sizes were made to allow better coverage of the WV. The smaller patterns are printed using a laser printer while the larger ones are drawn with a plotter. In both cases they are fixed on a planar hard surface. The circle centers projected on the image plane of a camera are automatically detected and matched with the actual calibration pattern using four specially colored reference circles detected in HSV space for more robustness. Using these four matches of the planar pattern, a homography is estimated to match the remaining black circles. It is worth noting that the four colored reference circles are of different configurations for scaled versions of the calibration target, allowing automatic multiple target identification in the same procedure. Typically, two target sizes are used.

Sampling pyramid In order to estimate the intrinsic camera parameters (including distortion), it is important to adequately sample the whole WV of each camera. In a fixed environment, the method presented in [8] is used to sample the WV by moving a calibration bar to known 3D positions with calibrated tripods. This method is not flexible enough for a transportable setup. Instead, a semi-automatic procedure has been developed to assure a good coverage of the WV without having to move the calibration pattern to known 3D positions. First, an initial estimate is needed for the intrinsic parameters in order to estimate the extent of the volume. This estimate can be the result of a previous calibration run, or can be derived from the lens manufacturer's specifications. The user then proceeds to manually show the calibration target to each camera at the near and far limits of its WV. The system continuously tracks the target and provides an audible feedback whenever it is correctly detected within an image frame. From the initial intrinsic parameters and the homography between the model target and its image, it is possible to compute the target position with respect to the camera and thus define an approximate sampling volume by integrating the nearest and farthest detected set of points.

Then, knowing the minimum and maximum depths of the view field, respectively $z_{\min}$ and $z_{\max}$, a sampling pyramid of n levels can be constructed, as illustrated in Figure 2. For each level $i = 0, 1, \ldots, n-1$ of this pyramid, $(i+1)^2$ sampling points are uniformly distributed on a spherical surface patch that lies within the view field, and is centered on the camera's center of projection, with radius $r_i = [(n-i-1)z_{\min} + iz_{\max}]/(n-1)$.

First step: intrinsic parameters calibration Given the sampling pyramid, the user again moves the calibration target within the WV until he covers all sampling points. A sample point is said to be covered whenever the distance between this sample point and the center of the target is sufficiently small. The system provides audible feedback each time the target covers a new sample point, and records the corresponding 2D positions of all circle centers in the image. When all points in the pyramid are covered, the system proceeds with the collected data to estimate both intrinsic and (backward) distortion parameters using Heikkilä's algorithm.

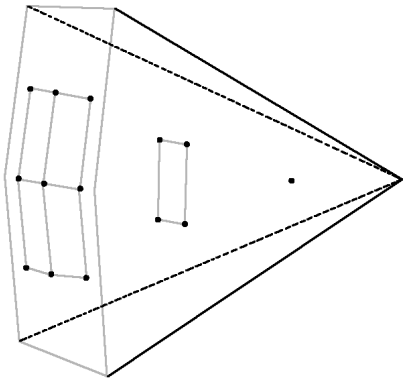


Figure 2: Example of a 3-level sampling pyramid.

Second step: reference frame and optimization In the second stage of the calibration procedure, the global reference frame is set to the reference frame of the first camera. The user simply moves the target around in the WV while the system tries to simultaneously detect it in all c images. Data is collected whenever the target is detected in at least two images. The system uses the same sampling pyramid defined previously as a guide to collect images. For each sample point of each camera, the system keeps a list of references to sets of detected targets. Therefore, the bookkeeping can be performed independently for each camera without duplicating the collected data. Each time a target is detected in a subset of cameras, $\{C\}$, the system computes the closest sample point in the pyramid, for each camera in $\{C\}$. For each of these cameras, the list associated to the sample point is scanned and elements for which the set of cameras is included in $\{C\}$ are removed. Then, the new observation is either dismissed if $\{C\}$ is included in any element already in the list or it is inserted in the list.

When a new set of images is available, the system uses homography to compute the rigid transformation between each of the cameras where the target was detected. This gives an initial estimate of the extrinsic parameters for each camera. As soon as all camera positions are known in a common coordinate system, the user can end the data collection and proceed to the optimization. These positions can be estimated directly or indirectly through intermediate transformations when the target cannot be seen from all cameras at the same time.

The intrinsic parameters of all cameras and the $6(c-1)$ free extrinsic parameters are then optimized using the Levenberg-Marquardt Algorithm [6] to minimize either the registration error or the registered pattern projection error of all detected circle centers. The algorithm is initialized with the intrinsic parameters obtained from the first step and the extrinsic parameters estimated from homography.

3D reconstruction Matched image points in two or more images are used to estimate 3D points by simple triangulation. To make these estimates from multiple images, there are several possibilities [2]. We implemented the classical method that consists in first eliminating distortion on the observed matched image points, and computing a least-square estimate of the intersecting rays in 3D space.

3 Assessing the calibration quality

In this section, procedures for locally assessing the uncertainty of a 3D measurement are described.

Reconstruction error After having completed the calibration procedure, the local uncertainty of the system can be assessed by once again moving the calibration target throughout the WV. For each detected target, the circle centers are matched between cameras and their 3D position is computed. The recovered 3D structure of the set of target points is then matched to the actual target rigid structure by registration [4] to yield $\mathbf{T}$, the rigid transformation from an observed object $\{\hat{\mathbf{w}}\}$ to its model $\{\mathbf{w}\}$. The mean reconstruction error for q control points on the model is then given by:

$$\epsilon_r = \sqrt{\frac{1}{q}\sum_{i=1}^{q} ||\mathbf{w}_i - \mathbf{T}\hat{\mathbf{w}}_i||^2} \tag{1}$$

This error provides a quantitative estimation of the local coherence (rigidity) of the measurements.

Epipolar error From a pair of matched points $\mathbf{m}_l$ and $\mathbf{m}_r$ observed in cameras l and r and the fundamental matrix $\mathbf{F} = \mathbf{A}_r^{-1}\,[\mathbf{t}_{lr}]_\times\,\mathbf{R}_{lr}\mathbf{A}_l^{-1}$, the epipolar error in image r is given by:

$$\epsilon_e = \frac{|\tilde{\mathbf{m}}_l^T\mathbf{F}\tilde{\mathbf{m}}_r|}{\sqrt{\left(\tilde{\mathbf{m}}_l^T\mathbf{f}_1\right)^2 + \left(\tilde{\mathbf{m}}_l^T\mathbf{f}_2\right)^2}} \tag{2}$$

where $\tilde{\mathbf{m}}$ is the vector $\mathbf{m}$ augmented by adding 1 as the last element and $\mathbf{f}_i$ is the ith column of $\mathbf{F}$. This error for camera r and the corresponding 3D position are computed and memorized for each point of the pattern and for each camera l. Moving the target throughout the volume provides a quantitative estimation of the calibration quality for camera r.

4 Experiments

4.1 Hardware and synchronization

The current experimental system encompasses $c = 4$ progressive scan digital color cameras (Pulnix TMC-6700-CL, 60 frames/s) mounted on tripods, typically positioned in the corners of the working area or on a circular

arc layout. Each camera is connected to a distinct standard PC through a CameraLink interface board (Matrox Meteor II / CameraLink). This distributed architecture was devised in order to maximize throughput for real-time processing, by conducting basic 2D image processing operations independently on each image stream. Extracted information and features are then routed to a Beowulf cluster of computers for high-level processing and fusion of information between image streams. This approach benefits from both the low cost and high performance of current PC hardware.

But distributing the image streams on different computers creates synchronization problems that must be dealt with adequately. First, the cameras are genlocked to ensure that all frames are generated simultaneously. A master synchronization board was also designed to dispatch a *vsync* signal to all cameras, and to provide basic handshake between the PCs. Before starting the processing of a new image, each PC waits for a ready signal from the others. In case of heavy computational burden on one or more of the PCs, the others stay synchronized simply by skipping frames until everyone is ready to go on with the next acquisition. In order to integrate local results, each host computes a globally consistent time stamp using its internal hardware timer combined with a network-based time barrier that is enforced periodically. This barrier is necessary because the operating system that we use (Windows 2000) provides no real-time guarantees, and sometimes causes the images to get out of sync when it decides unexpectedly to do house cleaning chores during critical timing operations.

To validate this synchronization method, we have placed a *vsync* triggered electronic digital counter in front of the four cameras, so that the counter value corresponds to a frame number. Then, using OCR software, we were able to compare the recognized frame numbers for equal time stamps in order to verify the effective synchronization of the computers. The time barrier was enforced every 1000 frames, and the experiment lasted over 15 consecutive hours. During that time, 99.83% of the frames matched perfectly, 0.15% were rejected because of OCR errors, and only 0.014% were definitely wrong because of a single machine that lost synchronization. These out-of-sync frames, however, were all within the same frame bloc and synchronization was automatically restored by the subsequent network time barrier.

4.2 Results

Examples of calibration results are provided for three different camera configurations. For the first configuration (Figure 3), the cameras are arranged roughly in the four corners of a 5×5 m rectangle. In the second configuration (Figure 4), the cameras are positioned more or less along a circular arc. In the third configuration, (Figure 5), the cameras are placed on the corners of a vertical cross.

To estimate the intrinsic parameters, we typically use a sampling pyramid constructed with $n = 3$ levels (14 sample points), which allows for good coverage of the WV. Using the four cameras, a residual error of less than 0.13 pixels is obtained after back projection of the detected circle centers. For the extrinsic parameters, using respectively 37 and 46 target positions, the first two configurations were calibrated by minimizing the registration error for all detected circle centers. The residual 3D registration errors were on average 3.3 mm and 2.7 mm. The third configuration was calibrated by minimizing the registered pattern projection error of all detected circle centers using 21 target positions. The average residual projection error was 3.4 pixels.

Reconstruction error Errors can be visualized by projecting on the floor the average errors for the corresponding vertical columns. Figures 3 to 5 illustrate the reconstruction error (ϵ_r) distribution within the observed area for the three configurations. The bullet diameters[1] represent the relative error amplitudes. In the first configuration (Figure 3), errors range from 1.3 mm to 10 mm with an average of 3.8 mm. These statistics stem from 3949 detected targets. For the second configuration (Figure 4), they range from 1.8 mm to 17 mm (1983 targets) with an average of 5.9 mm. For the third configuration (Figure 5), errors range from 1.6 mm to 14 mm (809 targets) with an average of 6 mm. The 12.7 mm CCD cameras were mounted with 6 mm lenses. The sizes of the calibration targets were 30 and 60 cm (Figure 1). The depth of the WV is approximately 3 m for the first configuration, versus 4 m for the second and the third. The larger observed errors for the arc and cross configurations are explained by the depth increase as well as the configuration where, in these cases, the cameras are positioned on one side of the WV.

Epipolar error Again, errors are visualized by projecting on the floor the average errors for the corresponding vertical columns. The third column of Figure 6 illustrates the epipolar error (ϵ_e) distribution within the WV for the cross configuration. The gray circle in each field of view is the approximate position of an observed human; it is discussed in the *Tracking people* paragraph. The mean error is 5.0, 3.4, 2.9 and 2.3 pixels for cameras 0, 1, 2 and 3, respectively. These statistics stem from 615, 525, 448 and 387 detected targets for each camera. The largest shown errors are 50, 25, 24 and 9.9 pixels, respectively. Comparing these results with Figure 5 suggests that the reconstruction error provides incomplete information about the calibration quality. For instance, the epipolar error near cameras 0, 1 and 2 is much larger than the reconstruction

[1] Please note that error bullets are much enlarged.

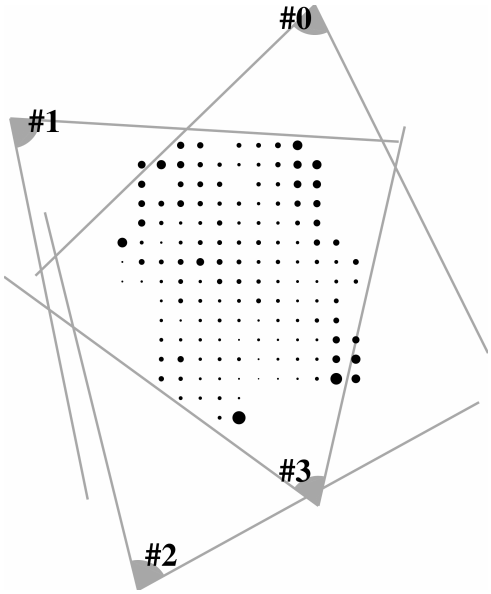


Figure 3: Working volume and field of view of the cameras for the square configuration. The bullet diameters represent the reconstruction error amplitudes vs. position.

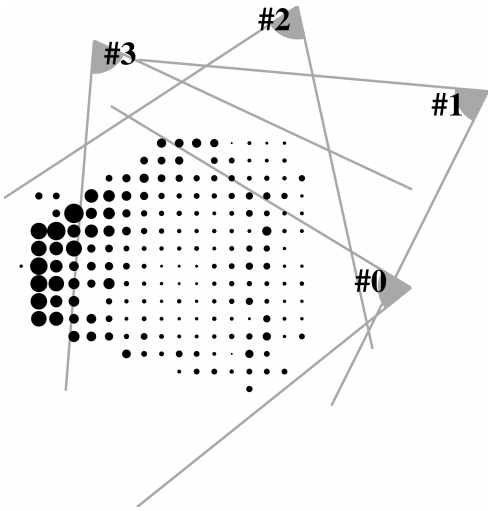


Figure 4: Working volume and field of view of the cameras for the arc configuration. The bullet diameters represent the reconstruction error amplitudes vs. position.

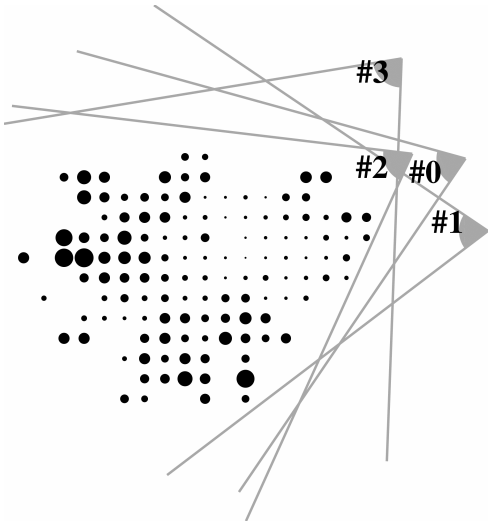


Figure 5: Working volume and field of view of the cameras for the cross configuration. The bullet diameters represent the reconstruction error amplitudes vs. position.

error would lead to believe. Nevertheless, the average error is larger for camera 0 than for the other three, which indicates that the measurements obtained from this camera are less accurate.

Sampling pyramid To demonstrate the need of adequate sampling of the WV, the cross configuration (Figure 5) was calibrated with partial sampling pyramids. The estimated z_{min} and z_{max} were used to cover only parts of the original WV's depth using $z'_{min} = z_{min} + d_{min}(z_{max} - z_{min})$ and $z'_{max} = z_{min} + d_{max}(z_{max} - z_{min})$ where $0 \leq d_{min} \leq d_{max} \leq 1$. Three sampling pyramids were defined to cover the near, center and far WV: *near* with $d_{min} = 0$ and $d_{max} = 0.2$, *center* with $d_{min} = 0.4$ and $d_{max} = 0.6$ and *far* with $d_{min} = 0.8$ and $d_{max} = 1$. In all cases, the configuration was calibrated by minimizing the registered pattern projection error of all detected circle centers. For *near*, the residual error of 17 target positions was 3.9 pixels, for *center*, the residual error was 2.4 pixels (41 targets) and for *far*, the residual error was 3.0 pixels (62 targets).

The quality of the calibration parameters was then evaluated by estimating the reconstruction error of 809 detected targets in the complete WV. The distribution of the reconstruction error in the WV is shown in Figure 7. The reconstruction error for *near* ranges from 3.9 mm to 53 mm with an average of 32 mm. For *center*, it ranges from 1.9 mm to 26 mm (average of 13 mm) and for *far*, it ranges from 2.0 mm to 43 mm (average of 7.6 mm). As can be seen from Figure 7, the spatial error distribution is strongly correlated with the sampling pyramid location. The parameters obtained with the *near* pyra-

mid can be used to compute adequate reconstructions in the near WV but are completely useless in the far WV (Figure 7(a)). The same observation holds for the *center* pyramid (Figure 7(b)) and, to a lesser extent, for the *far* pyramid (Figure 7(c)). In all cases, the parameters obtained with the full pyramid (Figure 5) give a globally better 3D reconstruction. However, the local reconstruction error is smaller with the partial pyramids. For example, the *far* pyramid gives a better reconstruction in the far WV than the full pyramid. It is thus advantageous to adequately sample the *intended* WV for a specific application. The proposed sampling pyramid allows this sampling to be performed in a flexible manner without the need for known landmarks in the environment.

Tracking people Figures 6 and 8 show snapshots of a walking human in each of the three configurations, as seen by the four cameras at a given instant.

Markers were automatically detected in Figure 6 and corresponding 3D coordinates were computed and back projected into each image. The position of the subject in the WV of the cameras is represented by gray circles in the third column of Figure 6. As can be seen from this figure, the subject's position in the WV of cameras 0 and 1 (Figures 6(c) and 6(f)) overlaps a zone of larger epipolar error than in the WV of the other two cameras. This large epipolar error is observed on the projected features; in cameras 0 and 1 (Figures 6(b) and 6(e)), the projected 3D model lies about 10 pixels away from the segmented markers while it lies less than 3 pixels away in the other cameras (Figures 6(h) and 6(k)). The estimated epipolar error is thus a useful information for weighting the contributions of multiple images in order to measure objects. In this case, cameras 0 and 1 would contribute less than the other in this portion of the WV.

Matching feature points in Figure 8 were hand picked and corresponding 3D coordinates were computed and then back projected into each view to illustrate the quality of the calibration (one point is on the nose tip). The calibration results obtained in the *Reconstruction error* paragraph (Figures 3 and 4) indicate that the maximum deformation for an object of the size of a human torso is about 1 cm anywhere within the volume.

5 Conclusion

A procedure for calibrating a reconfigurable network of cameras coupled to a real-time acquisition and processing network of computers was described. The calibration procedure begins with a separate estimation of the intrinsic camera parameters. Next, the relative positions of the cameras are estimated, and then all parameters are globally optimized. The procedure is interactive both for intrinsic and extrinsic calibration as the system assists the sampling of the working volume. Once the calibration is complete, new images are gathered to provide independent observations in order to assess the local uncertainty of the camera system. The epipolar error is available to weight the measure of each camera. It is also possible to review the system configuration when the 3D reconstruction error requirements are not met.

In the future, non planar portable targets will be exploited to facilitate the acquisition of images across multiple, possibly orthogonal, viewpoints. Due to strong incident angle, such a configuration cannot be easility calibrated with standard pairwise calibration techniques.

Acknowledgements: This work is supported by NSERC Canada through scholarships to S. Drouin and R. Poulin and research grants to P. Hébert and M. Parizeau. The hardware for this project was funded through the Canadian Funds for Innovation (CFI).

References

[1] X. Chen, J. Davis, and P. Slusallek. Wide area camera calibration using virtual calibration objects. In *CVPR00*, pages II:520–527, 2000.

[2] R. Hartley and P. Sturm. Triangulation. *Computer Vision and Image Understanding*, 68(2):146–157, 1997.

[3] J. Heikkilä. Geometric camera calibration using circular control points. *PAMI*, 22(10):1066–1077, October 2000.

[4] B.K. Horn. Close-form solution of absolute orientation using quaternions. *Journal of the Optical Society of America*, 5(7):1127–1135, 1988.

[5] T. Kanade, H. Saito, and S. Vedula. The 3D room: Digitizing time-varying 3D events by synchronized multiple video streams. CMU-RI-TR-98-34, Carnegie Mellon University, 1998.

[6] J.J. Moré. *The Levenberg-Marquardt Algorithm: Implementation and Theory*, volume 630 of *Lecture Notes in Mathematics*, pages 105–116. Springer Verlag, 1977.

[7] J.-C. Nebel, F. J. Rodriguez-Miguel, and W. P. Cockshott. Stroboscopic stereo rangefinder. In *Third Int. Conference on 3-D Digital Imaging and Modeling*, pages 59–64, 2001.

[8] P. Rander. *A Multi-Camera Method for 3D Digitization of Dynamic, Real-World Events*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, May 1998.

[9] Z. Zhang. A flexible new technique for camera calibration. *PAMI*, 22(11):1330–1334, November 2000.

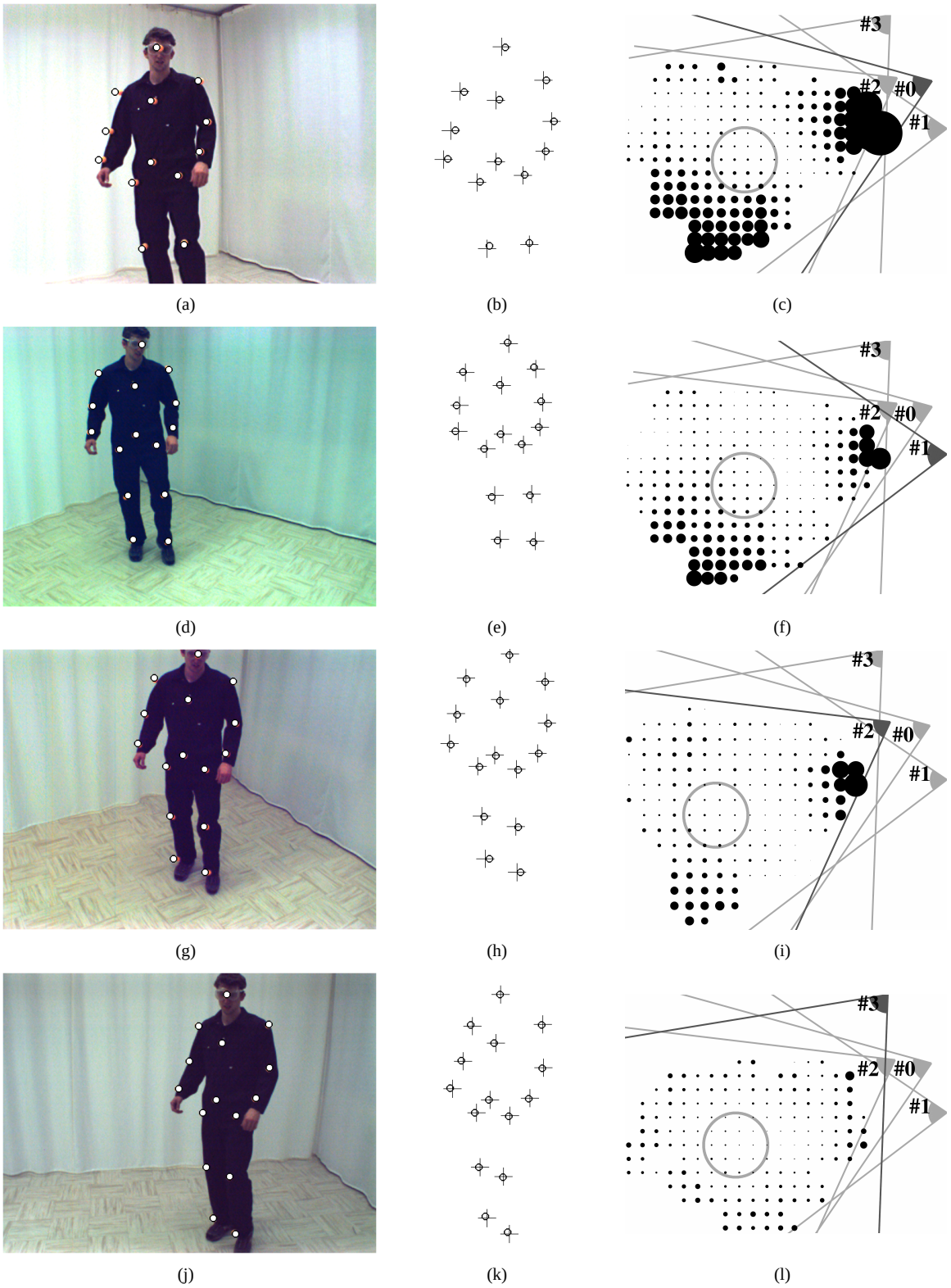
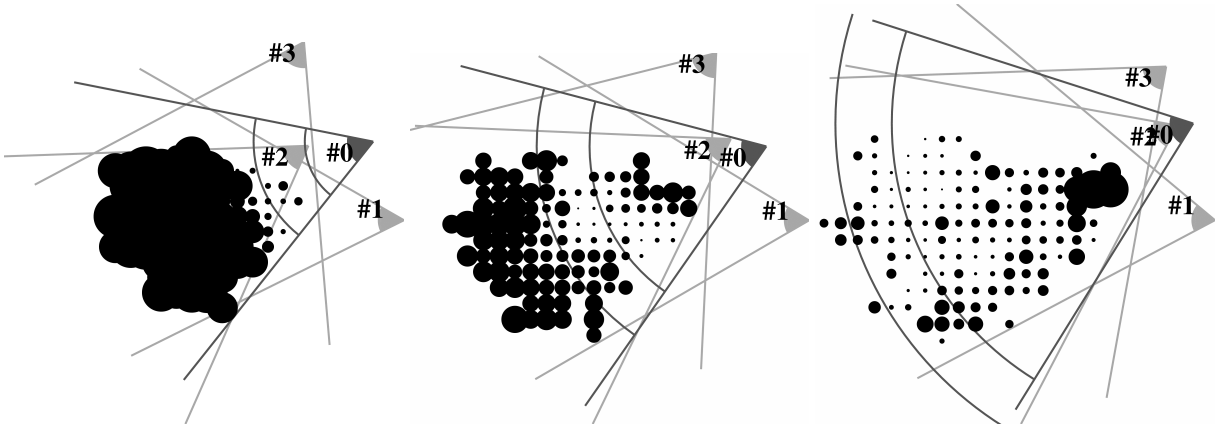
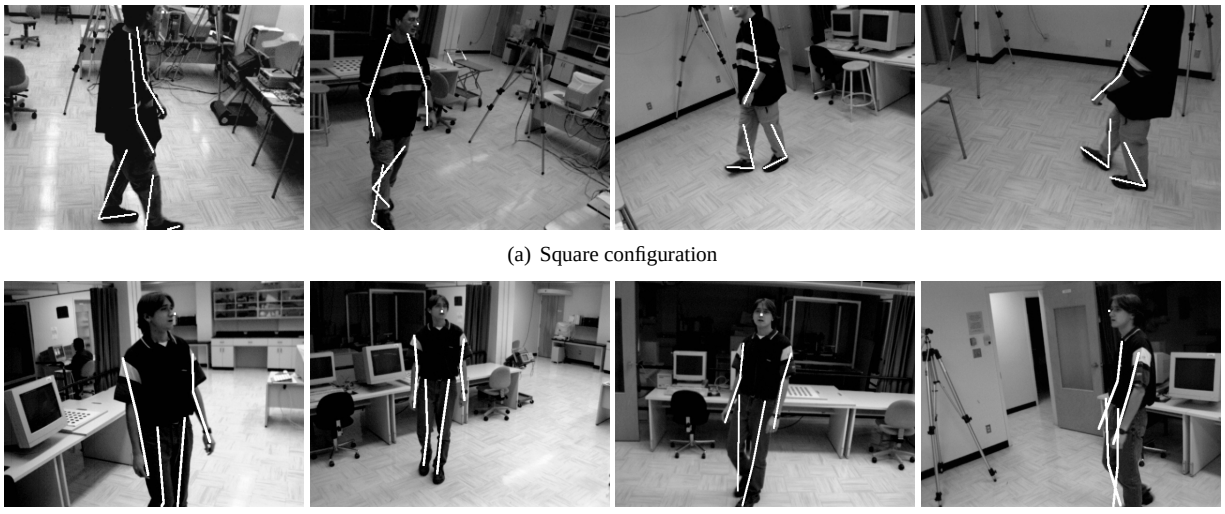


Figure 6: Left: reconstructed 3D features of a human projected into the input images for the cross configuration. Center: segmented (circle) and reconstructed (crosses) features. Right: epipolar error vs position where the camera of interest is highlighted. The gray circle is the position of the subject. From top to bottom: camera 0 to camera 3.



(a) Near working volume (b) Center working volume (c) Far working volume

Figure 7: Reconstruction error distribution for the cross configuration when the sampling pyramid is limited to the near, center or far working volume. The calibrated working volume of camera 0 is shown (arcs).



(a) Square configuration

(b) Arc configuration

Figure 8: Reconstructed 3D features of a human projected into the input images.