# Self Localizing Smart Camera Networks and their Applications to 3D Modeling

Camillo J. Taylor
Department of Computer and Information Science
University of Pennsylvania
cjtaylor@cis.upenn.edu

Babak Shirmohammadi
Department of Computer and Information Science
University of Pennsylvania
babaks@grasp.cis.upenn.edu

## Abstract

This paper describes a technology for localizing networks of embedded cameras and sensors. In this scheme the cameras and the nodes are equipped with controllable light sources (either visible or infrared) which are used for signaling. Each camera node can then automatically determine the bearing to all the nodes that are visible from its vantage point. From these angular measurements, the camera nodes are able to determine the relative positions and orientations of other nodes in the network.

The method is dual to other network localization techniques in that it uses angular measurements derived from images rather than range measurements derived from time of flight or signal attenuation. The scheme can be implemented with commonly available components and scales well since the localization calculations only require limited local communication. Further, the method provides estimates of camera orientation which cannot be determined solely from range measurements.

The localization technology can serve as a basic capability on which higher level applications can be built. The method could be used to automatically survey the locations of sensors of interest, to implement distributed surveillance systems or to analyze the structure of a scene based on the images obtained from multiple registered vantage points.

## 1 Introduction

As the prices of cameras and computing elements continue to fall, it has become increasingly attractive to consider the deployment of smart camera networks. Such camera networks could be used to support a wide variety of applications including environmental modeling, 3D model construction and surveillance.[3, 2, 7, 4]

One critical problem that must be addressed before such systems can be realized is the issue of localization. That is, in order to take full advantage of the images gathered from multiple vantage points it is helpful to know how the cameras in the scene are positioned and oriented with respect to each other.

In this paper we describe a deployment scheme where each of the smart cameras is equipped with a co-located controllable light source which it can use to signal other smart cameras in the vicinity. By analyzing the images that it acquires over time, each smart camera is able to locate and identify other nodes in the scene. This arrangement makes it possible to directly determine the epipolar geometry of the camera system from image measurements and, hence, provides a means for recovering the relative positions and orientations of the smart camera nodes.

A number of approaches to recovering the relative positions of a set of cameras based on tracked objects have been proposed in the literature [1, 6]. These approaches can be very effective in situations where one can gather sufficient correspondences over time. In contrast, the approach proposed here directly instruments the sensors and provides rapid estimates of the sensor field configuration using relatively modest computational and communication resources.

## 2 Implementation

Figure 1 diagrams the basic elements of our vision based localization system. Here we show a small network of 3 nodes, two of which are equipped with cameras. We will begin by discussing how localization proceeds in this simple case and then describe how the scheme can be extended to handle multiple nodes.

In the first stage of the localization process, the nodes signal their presence by blinking their lights in a preset pattern. That is, each of the nodes would be assigned a unique string representing a blink pattern such as 10110101, the node would then turn its light on or off in the manner prescribed by its string. Similar temporal coding schemes are employed in laser target designators and freespace optical communication schemes. These blink patterns provide a means for each of the camera equipped nodes to locate other nodes in their images. They do this by collecting a sequence of images over time and analyzing the image intensity arrays to locate pixels whose intensity varies in an appropriate manner. This approach offers a number of important advantages, firstly it allows the node to both localize and uniquely identify neighboring nodes since the blink patterns are individualized. Secondly, it allows the system to reliably detect nodes that subtend only a few pixels in an image which provides an avenue for further miniaturization of the smart camera nodes.

Figure 2 shows the results of the blinker detection phase on a typical image. Here the detected locations in the image are labeled with the unique codes that the system found.

Once the nodes have been detected and localized in the images, we can derive the unit vectors, $v_{ab}$, $v_{ac}$, $v_{ba}$ and $v_{bc}$ that relate the nodes as shown in Figure 3. Here we assume that the intrinsic parameters of each of the cameras (focal

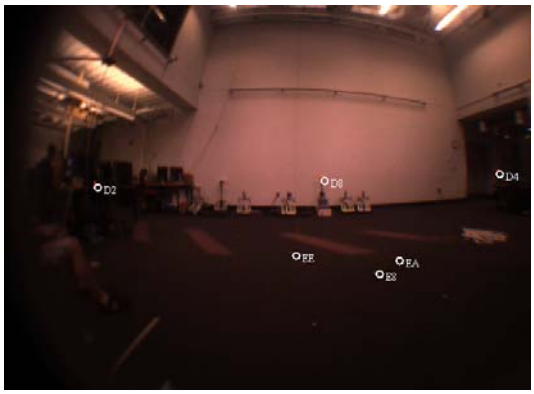
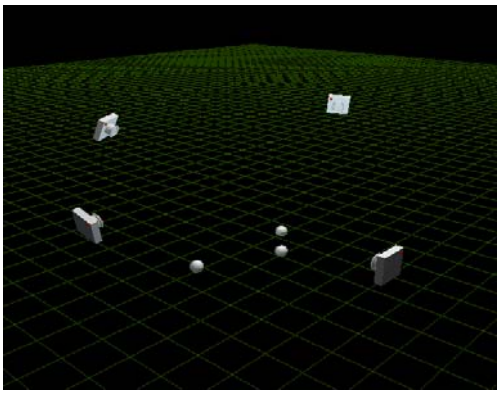
a.  b. 

**Figure 2. This figure shows the results of automatically localizing a constellation of 4 smart cameras and 3 blinker nodes. The image obtained from one of the smart cameras is shown in a while the localization results are shown in b.**

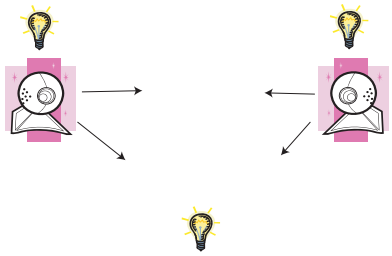


**Figure 1. This figure shows the basic elements of the proposed localization scheme. It depicts two smart camera nodes equipped with controllable light sources and a third blinker node.**

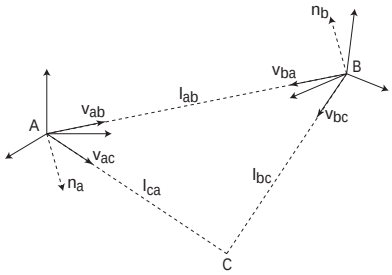



**Figure 3. This figure depicts the relative vectors and lengths in our 3 node localization problem.**

length, principal point, distortion coefficients) have been determined in an offline calibration step. These parameters allow us to relate locations in the image to direction vectors in space.

From the vectors $v_{ab}$, $v_{ac}$, $v_{ba}$ and $v_{bc}$, we can derive two additional vectors $n_a$ and $n_b$ which represent normalized versions of $(v_{ab} \times v_{ac})$ and $(v_{ba} \times v_{bc})$. These vectors correspond to the normal to the plane containing the three nodes, A, B and C expressed with respect to frames $A$ and $B$ respectively. Here we note that the unit vectors $v_{ab}$, $v_{ba}$, $n_a$ and $n_b$ are related by a rotation matrix $R_{ab}$ which captures the relative orientation of camera frames A and B.

$$v_{ab} = -R_{ab} v_{ba} \quad (1)$$

$$n_a = -R_{ab} n_b \quad (2)$$

From the two perpendicular unit vectors, $v_{ab}$ and $n_a$, we can construct the orthonormal matrix $R_a \in SO(3)$ as follows: $R_a = [\ v_{ab} \quad n_a \quad (v_{ab} \times n_a)\ ]$ Similarly, from the orthogonal unit vectors $-v_{ba}$ and $-n_b$ we construct the matrix $R_b = [\ -v_{ba} \quad -n_b \quad (v_{ba} \times n_b)\ ]$

From equations 1 and 2 we deduce that:

$$R_a = R_{ab} R_b \quad (3)$$

which in turn yields the following expression for $R_{ab}$:

$$R_{ab} = R_a (R_b)^T \quad (4)$$

Once we have ascertained the relative orientation of the two cameras. We can recover the relative position of the three nodes by considering the following homogenous linear system.

$$l_{ab} v_{ab} + l_{bc} (R_{ab} v_{bc}) - l_{ca} v_{ac} = 0 \quad (5)$$

Here the unknown variables $l_{ab}$, $l_{bc}$ and $l_{ca}$ denote the lengths of the segments AB, BC and CA. Since this system is homogenous we can only resolve the configuration of the nodes up to a positive scale factor.

The scheme that we have described essentially corresponds to the calibration of a stereo system where both of the epipoles can be located and measured directly in the imagery. In this configuration, we require only a single additional point to resolve the relationship between the two camera frames. Since the epipoles are directly measured, the localization scheme is quite stable numerically and can be expected to yield accurate results as long as we avoid the singular configuration where all three nodes are collinear.

Larger networks of smart cameras and sensors can be localized by considering the relationship between triangular subgraphs of the visibility graph as shown in Figure 4. In this graph, the directed edges indicate that a particular smart camera can view another node in the graph. The triangles indicate triples of nodes that can be localized using the scheme

described previously. The localization results from triangles that share an edge can be fused together into a common frame of reference. Therefore, if the set of localization triangles is fully connected, the entire network can be fully localized. Alternatively, by analyzing the connected components in the induced graph of localization triangles, one can automatically determine which sets of cameras can be localized to a common frame. The entire localization procedure is capable of determining the relative location and orientation of the nodes up to a scale factor, this scale can be resolved by measuring the distance between any pair of nodes. Figure 6 shows the final result of localizing a constellation of four smart cameras and three blinker nodes. Note that we do *not* require all of the nodes to have cameras - once we have localized two or more of the cameras we can localize other nodes equipped with lights through simple triangulation. This is an important advantage since it means that we can deploy a few smart camera nodes in an environment and use them to localize other smaller, cheaper sensor motes that are simply outfitted with blinkers.

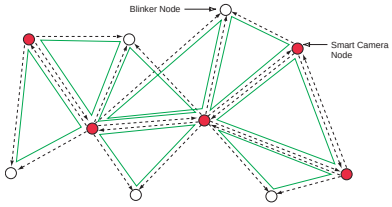



**Figure 4. Larger networks of smart cameras and sensors can be localized by considering the relationship between triangular subgraphs of the visibility graph.**

It is important to note that in this framework angular measurements derived from images and range measurements derived from other sources are treated as *c*omplementary sources of information. Measurements derived from the vision system can be used to determine the relative orientations of the camera systems which is important information that cannot be derived solely from range measurements. On the other hand, range measurements can be used to resolve the scale ambiguity inherent in angle only localization schemes. Similarly angular measurements can be used to disambiguate the mirror reflection ambiguities that are inherent in range only localization schemes. Ultimately it is envisioned that smart camera networks would incorporate range measurements derived from sources like the MIT Cricket system or Ultra Wide Band radio transceivers. These measurements could be used to improve the results of the localization procedure and to localize nodes that may not be visible to the smart camera nodes.

## 2.1 Refining Pose Estimates

The previous section described how neighboring cameras can compute their relative position and orientation based on corresponding image measurements. This process can be done in a completely decentralized manner using only local communication and will produce accurate relative location estimates which is what is typically what is required to fuse measurements from neighboring sensors.

If necessary, the estimates for node position and orientation produced by this process can be further refined by a scheme which takes account of all available measurements simultaneously. In this refinement step the localization process is recast as an optimization problem where the objective is to minimize the discrepancy between the observed image measurements and the measurements that would be predicted based on the estimate for the relative positions and orientations of the sensors and cameras. This process is referred to as Bundle Adjustment in the computer vision and photogrammetry literature.

In the sequel we will let $u_{ij} \in \mathbb{R}^3$ denote the unit vector corresponding to the measurement for the bearing of sensor $j$ with respect to camera $i$. This measurement is assumed to be corrupted with noise. The vector $v_{ij} \in \mathbb{R}^3$ corresponds to the predicted value for this direction vector based on the current estimates for the positions and orientations of the sensors. This vector can be calculated as follows:

$$v_{ij} = R_i(T_j - T_i) \tag{6}$$

In this expression $R_i \in SO(3)$ denotes the rotation matrix associated with camera $i$ while $T_i, T_j \in \mathbb{R}^3$ denote the positions of camera $i$ and sensor $j$ respectively (note that sensor j could be another camera).

The goal then is to select the camera rotations and sensor positions so as to minimize the discrepancy between the vectors $u_{ij}$ and $v_{ij}$ for every available measurement. In equation 7 this discrepancy is captured by the objective function $O(\mathbf{x})$ where $\mathbf{x}$ denotes a vector consisting of all of the rotation and translation parameters that are being estimated.

$$O(\mathbf{x}) = \sum_{\mathbf{i},\mathbf{j}} \|\mathbf{u_{ij}} - \frac{\mathbf{v_{ij}}}{\|\mathbf{v_{ij}}\|}\|^2 \tag{7}$$

Problems of this sort can be solved very effectively using variants of Newton's method. In these schemes the objective function is locally approximated by a quadratic form constructed from the Jacobian and Hessian of the objective function

$$O(\mathbf{x}+\delta\mathbf{x}) \approx O(\mathbf{x}) + (\nabla O(\mathbf{x}))^{\mathbf{T}}\delta\mathbf{x} + \frac{\mathbf{1}}{\mathbf{2}}\delta\mathbf{x}^{\mathbf{T}}(\nabla^{\mathbf{2}}O(\mathbf{x}))\delta\mathbf{x} \tag{8}$$

At each step of the Newton algorithm we attempt to find a step parameter space $\delta\mathbf{x}$ that will minimize the overall objective function by solving a linear equation of the form.

$$\delta\mathbf{x} = -(\nabla^2 O(\mathbf{x}))(\nabla O(\mathbf{x})) \tag{9}$$

Here we can take advantage of the fact that the linear system described in equation 9 is typically quite sparse More specifically, the Hessian matrix $\nabla^2 O$ will reflect the structure of the visibility graph of the sensor ensemble. This can be seen by noting that the variables corresponding to the positions of nodes $i$ and $j$ only interact in the objective function if node $i$ observes node $j$ or vice versa. For most practical deployments, the visibility graph is very sparse since any given camera typically sees a relatively small number of

nodes. This means that the computational effort required to carry out the pose refinement step remains manageable even when we consider systems containing several hundred cameras and sensor nodes.

## 3 Applications of Smart Camera Networks

Self localizing smart camera networks can serve as an enabling technology for a wide range of higher level applications. Here we focus on two applications where the images from the camera systems are used to derive information about the geometric structure of the environment.

### 3.1 Visual Hull Reconstruction

Multi camera systems are commonly used to derive information about the three dimensional structure of a scene. One approach to the reconstruction problem which is particularly well suited to the proposed self localizing smart camera network is the method of volume intersection which has been employed in various forms by a number of researchers [5]. This method can be used to detect and localize dynamic objects moving through the field of view of the smart camera network. Here a set of stationary cameras are used to observe one or more objects moving through the scene. Simple background subtraction is employed to delineate the portions of the images that correspond to the transient objects. Once this has been accomplished one can interrogate the occupancy of any point in the scene, P, by projecting it into each of the images in turn and determining whether or not it lies within the intersection of the swept regions. This process can be used to produce an approximation for the 3D structure of the transient objects by sampling points in the volume. The results of such an analysis are shown in Figure 5.

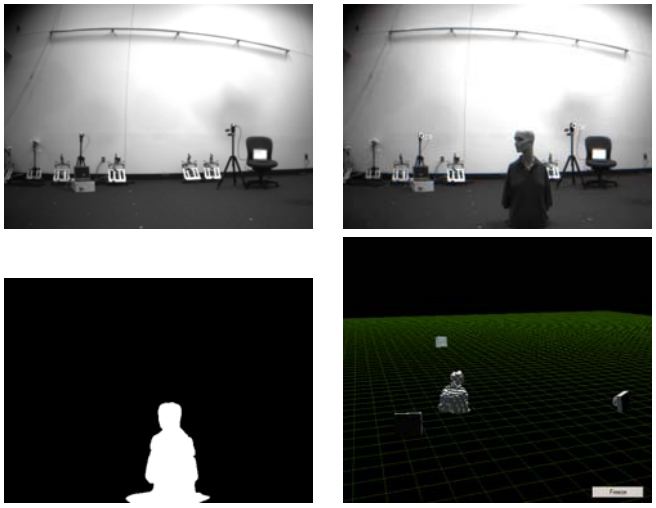


**Figure 5. (a) Background image of a scene (b) Image with object inserted (c) Results of the background subtraction operation (d) Results of applying the volumetric reconstruction procedure to the difference images derived from the three smart camera nodes**

In this application the ability to rapidly localize a set of widely separated cameras is a distinct advantage. Other implementations of this reconstruction scheme involve complex, time consuming calibration operations. This implementation, in contrast, could be be quickly deployed in an ad-hoc manner and would allow a user to localize and track moving objects such as people, cars or animals as they move through the scene.

### 3.2 Ad Hoc Range Finder

Another approach to reconstructing the 3D geometry of the scene using the imagery from the smart camera network involves establishing stereoscopic correspondences between points viewed in two or more images. If we are able to find such corresponding points we can readily reconstruct their 3D locations through triangulation. In order to employ this scheme we need a mechanism for establishing correspondences between pixels in one image and their mates in another.

One approach to establishing these inter frame correspondences is to employ structured illumination to help disambiguate the matching problem. This idea has been employed successfully in a number of stereo reconstruction systems. One such structured illumination scheme is depicted in Figure 7 where a projection system sweeps a beam of light across the surface of the scene. Correspondences can then be established by simply observing when various pixels in the two images are lit by the passing beam.

Figure 6 shows a pair of images acquired using such a structured light correspondence scheme. Here a plane of laser light is swept across the scene and the curves corresponding to the illuminated pixels in the two images are recovered. In each image, every point on the curve corresponds to a ray in space emanating from that camera position. To find the correspondence for that point in the other image we first project that ray into the other image to construct the corresponding epipolar line and then search along that line to find the corresponding pixel that is also illuminated by the laser plane as shown in Figure 7.

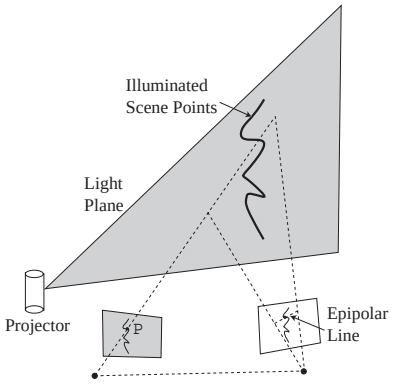



**Figure 7. At every point in time the projector illuminates a set of scene points along a planar curve in the scene. For every point on the projected curve in one image we can locate its correspondent in the other image by searching along the epipolar line in the other image.**

After sweeping the plane over the entire scene we are able to determine the range to most of the points in the scene that are visible from both camera positions even though those two camera positions are widely separated. Such a range map is shown in Figure 6c. This range scan was constructed by

**Figure 6. ((a) and (b) show Two images of a scene illuminated with a plane of laser light which is used to establish correspondences between the two views (c) shows the range map constructed based on the correspondences derived from a sequence of such images.**

sweeping the laser plane through 180 degrees in 1 degree increments.

It is important to note here that this range map is constructed in an ad-hoc manner since the relative positions and orientations of the cameras are reconstructed automatically using the self localization algorithm and the position and orientation of the projector are not needed to recover the scene depths. The proposed reconstruction scheme is interesting because it provides a mechanism for recovering the structure of an extended scene using an ensemble of small, cheap image sensors and beam projectors which can be deployed in an ad-hoc manner. This is in contrast to the traditional approach of recovering scene structure using expensive range sensors which must be carefully calibrated and aligned.

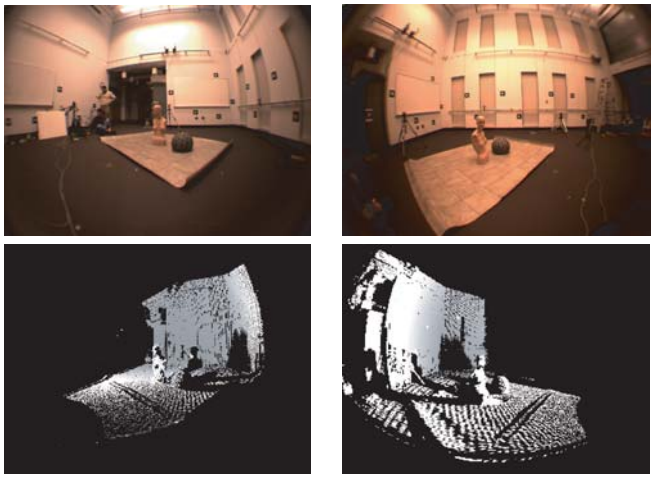


**Figure 8. In this experiment range maps of the scene were constructed from 4 different vantage points using different configurations of cameras and projectors. Two of these scans are shown here along with the corresponding images**

The scheme can be extended for use with multiple cameras and multiple beam projectors as shown in Figure 8. Here we are able to obtain multiple range maps of the scene taken from different vantage points using a collection of camera systems and projector positions. Importantly, since we are able to recover the relative positions of all of the cameras used here via the self localization scheme, all of the recovered range maps can be related to a single frame of reference. This provides an avenue to recovering the structure of extended environments by merging the range maps obtained from the different camera systems into a single coherent model of the scene.

## 4 Conclusions

This paper describes a scheme for determining the relative location and orientation of a set of smart camera nodes and sensor modules. The scheme is well suited for implementation on wireless sensor networks since the communication and computational requirements are quite minimal.

Self localization is a basic capability on which higher level applications can be built. For example, the scheme could be used to survey the location of other sensor motes enabling a range of location based sensor analyses such as sniper detection, chemical plume detection and target tracking. Further, the ability to automatically localize a set of smart cameras deployed in an ad-hoc manner allows us to apply a number of multi-camera 3D analysis techniques to recover aspects of the 3D geometry of a scene from the available imagery. Ultimately we envision being able to construct accurate 3D models of extended environments based on images acquired by a network of inexpensive smart camera systems.

## 5 References


[1] B. Dungan Ali Rahimi and Trevor Darrell. Simultaneous calibration and tracking with a network of non-overlapping sensors. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, 2004.

[2] Jason Campbell, Phillip B. Gibbons, Suman Nath, Padmanabhan Pillai, Srinivasan Seshan, and Rahul Sukthankar. Irisnet: an internet-scale architecture for multimedia sensors. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 81–88, New York, NY, USA, 2005. ACM Press.

[3] Wu chi Feng, Brian Code, Ed Kaiser, Mike Shea, Wu chang Feng, and Louis Bavoil. Panoptes: scalable low-power video sensor networking technologies. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 562–571. ACM Press, 2003.

[4] C. H. Lin, T. Lv, and I. B. Ozer W. Wolf. A peer-to-peer architecture for distributed real-time gesture recognition. In *International Conference on Multimedia and Exposition*, 2004.

[5] Wojciech Matusik, Christopher Buehler, Ramesh Raskarand Leonard McMillan, and Steven J. Gortler. Image-based visual hulls. In *SIGGRAPH*, 2000.

[6] M. Paskin S. Funiak, C. Guestrin and R. Sukthankar. Distributed localization of networked cameras. In *ISPN*, 2006.

[7] Z. Yue, L. Zhao, and R. Chellappa. View synthesis of articulating humans using visual hull. In *Proc. Intl. Conf. on Multimedia and Expo*, volume 1, pages 489–492, July 2003.