Deforming Objects Provide Better Camera Calibration

Ryan White
David Forsyth

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2005-3
http://www.eecs.berkeley.edu/Pubs/TechRpts/2005/EECS-2005-3.html

October 03, 2005

Deforming Objects Provide Better Camera Calibration

Ryan White, D.A. Forsyth
{ryanw,daf}@cs.berkeley.edu

October 3, 2005

Abstract

We present a method to calibrate perspective cameras from views of a deforming textured object. We proceed by finding an orthographic calibration, then enriching the camera model to include perspective effects and distortion terms. In the process, we establish the utility of using surface normals to calibrate cameras in the orthographic setting, with a proof that a metric reconstruction can be achieved in two orthographic views of two points combined with two normals. Our calibration object is a sheet of textured cloth. We show that a calibration applies to a specific region of space. Substantial improvements in reconstruction quality and in reprojection error can be obtained by using calibration objects that explore most of the relevant 3D volume. Using cloth as a calibration object allows easy calibration of large 3D volumes more accurately than typical planar calibration objects.

1 Overview

We present a new method of camera calibration targeting the reconstruction of cloth in moving sequences. Our requirements differ from common structure from motion estimation problems: we typically have a small number of cameras surrounding a roughly convex textured object, each point is typically viewed by three or fewer cameras, perspective effects are small but not negligible and neighboring points may be reconstructed from different cameras. As a result, small calibration errors result in large strains in the reconstructed cloth — a visually displeasing and physically implausible artifact.

Our main claim is that using a 'standard' fixed planar calibration object (a checkerboard in the bottom of the scene — see figure 9 in [14]) is not ideally suited for this application. First, this fixed object places unnecessary restrictions on the locations of the cameras. Even worse, when all cameras can view the same plane, this approach wastes resolution in the capture system. Second, we demonstrate that calibration is specific to a volume of space. Self calibration methods (such as [10]) obtain good reconstructions because they use the

same correspondences for both calibration and reconstruction. To achieve similar results from a secondary calibration pattern, one would need to move the calibration object through the volume before recording. However, it is difficult to guarantee that the appropriate volume in 3D is covered effectively. We opt for a more convenient technique: build the calibration into the deforming object, which then effectively explores the space. We calibrate over time — using the large numbers of correspondences to calibrate and thus guarantee that our calibration pattern covers the relevant volume.

It is not intuitive that one could calibrate cameras with a deforming object. In fact, as we show, the advantages of exploring the view volume are substantial. Our reprojection errors are significantly lower (figure 7) and, in contrast to other techniques, our reconstructions agree with physical cloth models. (figure 8)

Our calibration method differs from previous approaches. Observing that perspective effects are small, we calibrate in two steps: first, we compute an orthographic camera calibration using the printed cloth pattern, then 'enrich' the model to include perspective effects and distortion parameters. While general orthographic calibration of point clouds requires 3 views of a sufficient number of points to get a metric reconstruction, we prove in the appendix that using points and normals, a metric reconstruction can be obtained from only two orthographic views if two corresponding points and normals are known. While an orthographic view of any repeated pattern can provide normals [9], in this work we consider the case where we know the frontal texture pattern *a priori*.

2 Previous Work

Camera calibration is well understood, with comprehensive reviews in two excellent books [2, 6]. Software that implements the most common techniques is readily available on the Internet [1]. However, these tools have drawbacks when used to capture dynamic cloth, which moves through large volumes while showing minimal perspective effects. We begin with a review of the relevant terminology.

Camera parameters are described as a set of **extrinsic** (configuration) and **intrinsic** (focal length, camera center, etc.) variables. Camera calibration (determining the camera parameters) can be broken into two categories: **photogrammetric calibration**, where the geometry of the scene is known ahead of time to high precision; and **auto-calibration** where the structure of the scene is not known ahead of time and is simultaneously recovered from the 2D views.

The standard method involves: identifying **interest points**; using **appearance**, **epipolar** and **three view** constraints to build frame-frame correspondences between these points; obtaining a **projective reconstruction** — which yields geometry and cameras up to a 3D projective transformation — using one of several current factorization methods; and then using appropriate assumptions to obtain an **upgrade** to a Euclidean reconstruction. The reconstruction and cameras are then cleaned up with a **bundle adjustment**, which minimizes

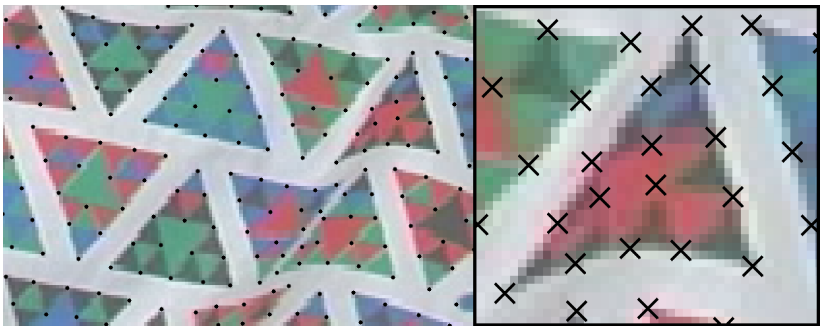


Figure 1: *We use a piece of cloth with a color coded pattern as our 'calibration object'. Our pattern contains information at multiple scales: the larger triangles have a color coded pattern that defines correspondence between views; the vertices of the smaller triangles provide large numbers of point correspondences. We extract this information in a coarse to fine search. Because we know the frontal pattern of the cloth and because we assume that there are few folds smaller than the scale of the smallest triangles, we can compute normals to each of the small triangles. The accuracy of rotation estimation in figure 3 confirms the validity of this assumption.*

reprojection error as a function of reconstruction and camera parameters.

While we do not directly assume the geometry of the scene, our work fits in with previous work in calibration assuming a calibration object. Early work in this area used a planar object with vertices at known locations [17]. More recently, improvements to the solution and readily available executables have made planar calibration commonplace [1, 19]. In contrast to these approaches, we make no assumptions about the scene geometry, but instead assume that normals can be computed from our fixed pattern. Our cloth pattern is composed of triangles at multiple scales and can be seen in figure 1.

Our application is cloth motion capture, which probably dates to [4], who mark surfaces with a grid and track the deformation of elements of this grid in the image. This work does not report a 3D reconstruction, because unlike the planar case correspondence is difficult with a periodic pattern. Guskov, Klibanov and Bryant give views of a 3D reconstruction, obtained by printing square elements with internal patterns on the surface, estimating local homographies at each element, then linking these estimates into a surface reconstruction [5]. The homographies tend to be noisy because perspective effects are weak or unobservable at the scale of an element, meaning that considerable work must be done to get a set of consistent estimates. [11, 12] use a calibrated stereo pair and SIFT feature matches to build a 3D model. [14] use a pseudo-random pattern of colored circles on custom-made garments to reconstruct both a parameterization and geometry. They use a checkerboard pattern painted on the floor of the studio to calibrate the cameras — constraining the location and motion of the cameras while consuming a large number of valuable pixels.

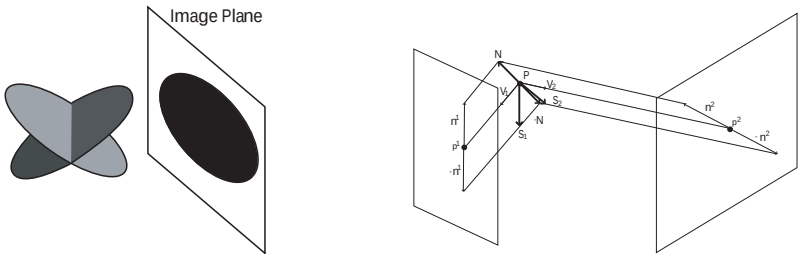


Figure 2: ***(Left)*** *Viewing a planar circle yields an ambiguous image of an ellipse in the image plane. Two possible circles correspond to the ellipse in the image.* ***(Right)*** *Notation for normal ambiguity in two views. There are two simple orthographic views of the point* $\boldsymbol{P}$*, with normal* $\boldsymbol{N}$*; view directions are* $\boldsymbol{V}_1$ *and* $\boldsymbol{V}_2$*. The text shows that* $\boldsymbol{S}_1$ *and* $\boldsymbol{S}_2$ *— ambiguous normals in their respective views — have heads lying on the same epipolar plane and that an incorrect match leads to a reconstruction of* $-\boldsymbol{N}$*.*

3 Normals and Orthography

Reconstruction from scaled orthographic views is now a standard algorithm (originating in [15]; many important variants appear in [6]). If there are more than two cameras, a metric reconstruction is available by enforcing scale and angle properties of the camera basis. However, this approach ignores knowledge of scene geometry (because we know what a triangle looks like, we can estimate a surface normal). A metric reconstruction isn't possible from two views in simple orthographic cameras without calibration of camera extrinsics or some known length or angle [8, 18]. Since the cloth moves fast and we may be stuck with only two views, and to incorporate our normal information, we adopt a method that exploits surface normals to obtain a metric reconstruction.

In a single scaled orthographic view, we know the normal of the plane on which the pattern element lies *up to a two-fold ambiguity* (e.g. [3, 9]). This ambiguity occurs because we can identify the cosine of the slant angle — usually written as $\cos\sigma$ — but not its value from a single view. For example, a scaled-orthographic view of a circle looks like an ellipse; we know the extent of the slant (and so the length of the normal) but the circle could have been slanted either forward or backward to yield this ellipse (figure 2). As a result, we know the projected normal up to an ambiguity of π radians.

The most natural way to incorporate this information into existing multiple view results is to think of the normal as an arrow of known length protruding from the surface at the point in question. The base of the arrow is the point in question, and projects as usual. The results above mean we know (up to a two-fold ambiguity) to what point in the image the head of the vector projects — in turn, having a normal from texture repetition is equivalent to having a second point *and* having some metric information *because we know the length of the normal vector.* For convenience, in what follows we refer to an isolated point

as a **point**, and a point with the normal information described as a **patch**.

3.1 The 3D Ambiguity of Normals

Assume that we are dealing with a pair of simple orthographic cameras. Furthermore, assume that the scale of the cameras is the same (we can obtain the relative scale from the size estimates for triangles), and that the extrinsics are calibrated. In a single view, the projected normal is known up to an ambiguity of π radians. What ambiguity is there in 3D reconstruction of the normal?

Write the normal as $\mathbf{N}$ and the i'th view vector pointing toward the camera (figure 2) as $\mathbf{V}_i$. In the i'th view, there are two possible 3D normals, $\mathbf{N}$ and $\mathbf{S}_i$ (the ambiguous normal in the i'th view). Because the image ambiguity is π radians, $\mathbf{N}$, $\mathbf{V}_i$ and $\mathbf{S}_i$ must be coplanar. Because the projected length of $\mathbf{S}_i$ is the same as the projected length of $\mathbf{N}$, $\mathbf{V}_i \cdot \mathbf{N} = \mathbf{V}_i \cdot \mathbf{S}_i$. This means that we have $\mathbf{S}_i = 2(\mathbf{N} \cdot \mathbf{V}_i)\mathbf{V}_i - \mathbf{N}$ The epipolar planes consist of every plane whose normal is $\mathbf{E} = \mathbf{V}_1 \times \mathbf{V}_2$. The "heads" of $\mathbf{S}_1$ and $\mathbf{S}_2$ lie on the same epipolar plane, because $\mathbf{E} \cdot \mathbf{S}_1 = \mathbf{E} \cdot \mathbf{S}_2 = -\mathbf{E} \cdot \mathbf{N}$. In the circumstances described, there are two possible matches for the "head" of the normal. First, the correct matches are available, resulting in a reconstruction of $\mathbf{N}$; second, one can match the image of the "head" of $\mathbf{S}_1$ with the image of the "head" of $\mathbf{S}_2$. The second case results in a reconstruction of $-\mathbf{N}$ (figure 2); this is easily dealt with, because visibility constraints mean that $-\mathbf{N} \cdot \mathbf{V}_i < 0$ for both i.

All this yields **Lemma:** *A metric reconstruction from two simple orthographic views is available from two patch correspondences. There is a maximum of sixteen ambiguous cases, yielding no more than four camera reconstructions.* **Proof:** (see appendix) There is an obvious **corollary:** *A fundamental matrix is available from two patch correspondences, up to at worst a four-fold ambiguity.*

3.2 Obtaining a Euclidean Camera Solution

To obtain a camera solution from two views, we perform a rough search over rotation matrices, then use gradient descent to refine the solution. Our cost function for this is based on the combination of two penalties: the reprojection cost of the points (in pixels) and the alignment cost of the normals (in degrees). Using $\mathbf{x}_i^j$ as the observed point i in view j, $\hat{\mathbf{x}}_i^j$ as the reprojected points, $\mathbf{N}_i^j$ as the respective normals and $\mathcal{R}$ as the rotation matrix between the two views, our costs can be computed as follows:

$$\begin{aligned} c_{\text{normals}} &= \textstyle\sum_i (1 - \mathbf{N}_i^1 \cdot \mathcal{R}\mathbf{N}_i^2) \qquad & c_{\text{pts}}^j &= \textstyle\sum_i \|\mathbf{x}_i^j - \hat{\mathbf{x}}_i^j\|^2 \\ c_{\text{pts}} &= c_{\text{pts}}^1 + c_{\text{pts}}^2 + \text{const} \cdot c_{\text{normals}} \end{aligned}$$

To obtain the reprojected reconstructed points $\hat{\mathbf{x}}_i^j$ from the observations and rotation matrix, we first move the center of gravity of the observed points to the origin. Then, we create an orthographic camera matrix $\mathcal{K}$ as $[\mathbf{i}^T, \mathbf{j}^T, \mathbf{r}_1^T, \mathbf{r}_2^T]$, where $\mathbf{i}, \mathbf{j}$ are in the coordinate axis directions as usual and $\mathbf{r}_1^T$ and $\mathbf{r}_2^T$ are the first two rows of the camera rotation matrix $\mathcal{R}$. Finally, we compute the reprojected

Figure 3: *Perspective effects of cloth are typically small. As a result, we can ignore them at first, obtain an orthographic reconstruction, then solve for a reconstruction from a perspective model. In these frames, we start with pairwise orthographic calibrations computed over a sequence of frames. Following the technique described in section 3.2, we minimize a combined objective that includes reprojection error and agreement in the rotated normals. To compute the agreement of the normals, we use the relative rotation matrix between two cameras to rotate normals from one viewpoint to another and report the angular alignment of the normals to verify the quality of the reconstruction. Averaging over 20 frames of a dynamic sequence viewed by three cameras, we have errors of 3.10 pixels and 1.48 degrees between the first view and the second and 2.96 pixels and 1.20 degrees between the second view and the third. While reprojection error is fair, the rotation error is very small. Because perspective effects are small on the scale of a triangle, the computed normals are very accurate.*

reconstructed points $\hat{\mathbf{x}}_i^j$, by using the pseudo-inverse of $\mathcal{K}$ to reconstruct and $\mathcal{K}$ to reproject. A grid search over rotation matrices $\mathcal{R}$ provides an estimate of the camera matrix and gradient descent refines the parametric camera model. Rotation estimates from this approach

Because normals are estimated from texture elements which effectively display no perspective effects individually (they are too small), our method yields a rotation estimate for *perspective* cameras from the locally valid assumption of scaled orthography. This remarkable fact is borne out by the excellent consistency of our normal estimates and our rotation estimates. In particular, applying our rotation to normal estimates between views yields alignment within two degrees. (Figure 3) The accuracy of this alignment also indicates that normals estimates themselves are accurate, even in images with significant wrinkles.

4 Perspective from Bundle Adjustment

At this point, we have a structure estimate and an estimate of camera extrinsics and scale assuming scaled orthography. Our cameras may not, in fact, be scaled orthographic cameras, and some lateral views of cloth display mild perspective effects (Figure 3). This results in potentially large reprojection errors

and poor reconstructions. We use the orthographic camera solutions as an initialization for a fuller perspective model. We then run bundle adjustment: a large minimization over the camera parameters (both intrinsic and extrinsic) and the reconstructed points. We refrain from a complete discussion of bundle adjustment here, and refer readers to [16] for more details.

4.1 An Orthographic Camera in a Perspective Model

To use the orthographic camera calibration as an initialization for the perspective model, we represent the orthographic camera in the richer perspective model.

Our camera model is based directly on the model adopted by [1], and is very similar to the models used in [19, 7]. While our model includes distortions due to lens artifacts, for cleanliness we drop these terms below. To distinguish between the orthographic and perspective equations, we adopt the subscript π for perspective and o for orthographic. Using (u, v) as coordinates in 2D, $(M = [X, Y, Z])$ as coordinates in 3D, $(\mathbf{R}, \mathbf{t})$ as rotation and scale, (u_0, v_0) as the principal point, (α, β) as perspective scale factors and s as the orthographic scale factor, we write the projection along a single camera axis as:

$$u^o = s(\mathbf{r}_1^o M + t_x^o) \qquad\qquad u^\pi = \alpha \frac{\mathbf{r}_1^\pi M + t_x^\pi}{\mathbf{r}_3^\pi M + t_z^\pi} + u_0^\pi$$

We note that the equations appear in a fairly similar form. By making the assignment $\alpha = t_z^\pi s$, in the limit of large t_z the perspective model becomes orthographic:

$$\lim_{t_z^\pi \to \inf} u_\pi = s(\mathbf{r}_1^\pi M + t_x^\pi) \left(\lim_{t_z^\pi \to \inf} \frac{t_z^\pi}{\mathbf{r}_3^\pi M + t_z^\pi} \right) + u_0^\pi = s(\mathbf{r}_1^\pi M + t_x^\pi) + u_0^\pi$$

Since the orthographic camera is the limit of the perspective camera, simple substitution allows us to use our orthographic calibration as an initialization for the perspective camera. We start by assigning the rotation matrices to be the same ($\mathbf{R}^\pi = \mathbf{R}^o$). An ambiguity exists in computing t_x^π and u_0^π from t_x^o. We assume that u_0^π should lie near the center of the camera, and use the following equations to complete the transition from orthographic to perspective:

$$t_z^\pi \to \inf \quad \begin{array}{lcllcllcl} \alpha & = & t_z^\pi s & u_0^\pi & = & \frac{\text{image width}}{2} & t_x^\pi & = & t_x^o - \frac{u_0^\pi}{s} \\ \beta & = & t_z^\pi s & v_0^\pi & = & \frac{\text{image height}}{2} & t_y^\pi & = & t_y^o - \frac{v_0^\pi}{s} \end{array}$$

4.2 Substitution in Practice

The only complication in this procedure is the initialization of the parameter t_z. As suggested in the previous section, we should initialize this to almost infinite value. However, in practice, we simply choose a value significantly larger than scene geometry suggests. Second, as outlined in the appendix, an orthographic view has an ambiguity in depth that does not occur in the perspective case.

Figure 4: *We demonstrate the ability to obtain a metric reconstruction from two orthographic views of a textured scene. We paste a triangle pattern on a box, forming two sections at a roughly 90 degree angle. Using both points and normals, we reconstruct the point locations by computing an orthographic camera calibration. Our calibration error is on average 2.41 pixels and 1.74 degrees. (photos taken with a consumer camera at maximum zoom to approximate orthography) To evaluate the results, we fit two planes to the two point sets and compute the angle between the planes to be* ***90.67 degrees*** *(we estimate 90 degrees from world geometry). When fitting the plane the MSE distance from the points to the plane is 2.57 pixels or 1.69 mm. The physical paper isn't completely flat — deviations of 2mm appear physically plausible.*

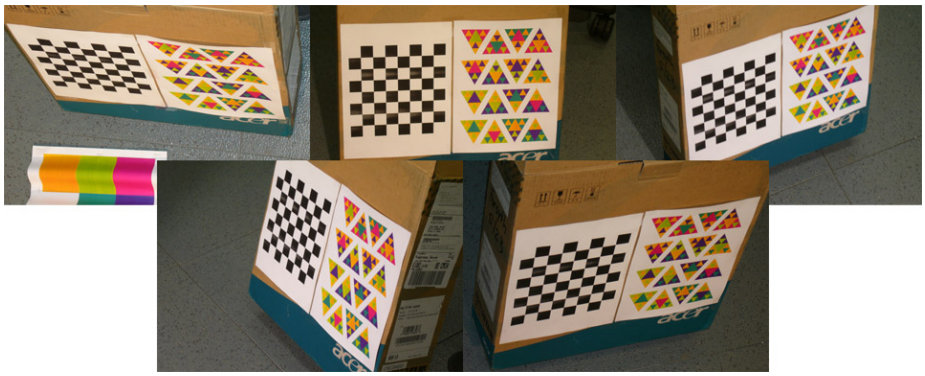
From an orthographic given camera, a flip in the z coordinate of the other camera matrices will produce a depth flipped reconstruction. However, in the perspective case, an erroneous flip in depth would cause nearby objects to appear smaller. To account for this, we search over the two cases for each camera matrix.

5 Experiments

We implemented the calibration framework established in this paper and printed several pieces of cloth and paper with our triangle pattern. Through experiments with real world data, we establish the following points:

A metric upgrade of a textured surface is available from two orthographic views. We establish this by photographing a scene of known geometry (two planes that form a right angle) and measure the angle from the reconstructed geometry to be within one degree. (Figure 4) Our rotation estimates are very accurate — implying that normal estimates are highly reliable.

Our calibration code for perspective scenes is comparable to existing methods in standard configurations. We show this by calibrating with a planar calibration object and calibration software available online. (Figure 5)



calib data (cam model)	evaluation data	reproj error
triangles (orthographic)	triangle pattern	0.995 pixels
triangles (perspective)	triangle pattern	0.318 pixels
triangles (perspective)	checkerboard	3.270 pixels
checkerboard (perspective)	checkerboard	0.172 pixels
checkerboard (perspective)	triangle pattern	1.220 pixels

Figure 5: *We compare our calibration pattern with standard software available in [1]. Our method does not perform as well, probably because it does not assume fixed geometry — limiting its accuracy in this confined case. Note that both methods generalize poorly, implying that calibration should be performed in the same region as reconstruction. As shown in figure 7, when calibrating larger regions of space, this problem becomes even more pronounced. Our calibration method starts by using texture cues to obtain a metric reconstruction for orthographic cameras, then 'enriches' the model to include perspective effects using bundle adjustment.*

5.1 Calibration of 3D volumes

Finally, we establish our main point: **calibration is specific to 3D volumes and calibration in the same volume is superior**. Empirically, we observe that calibration objects are better when they occupy the same 3D volume as the measured structure. To obtain better reconstructions, one needs to use a calibration object that is large and centrally located. However, in moving sequences, such calibration objects wastes resolution — occupying valuable pixels that could be used to estimate instead.

Using cloth as a calibration object in the same volume is substantially better than using a traditional calibration object. To gauge the effectiveness of our method we calibrate on one portion of the sequence, and compute errors on another portion of the sequence. However, in practice, calibration should be performed using the entire sequence.

In Figure 6, we give example images from a sequence of moving cloth and in Figure 7, we show that reprojection errors are significantly worse when using

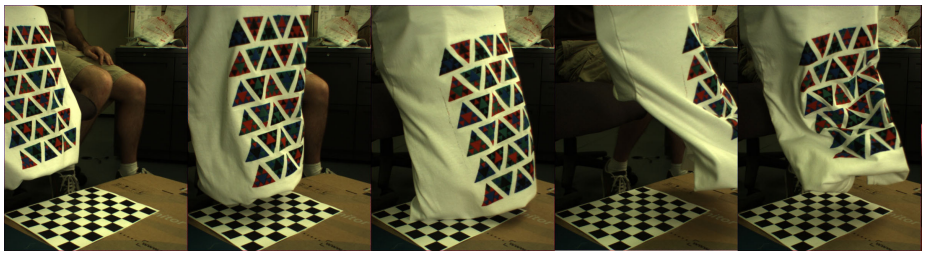


Figure 6: *Above is a selection of images taken from a single camera — the cloth moves significantly during different parts of the sequence. A planar calibration pattern (checkerboard) allows us to compare calibration methods (checkerboard calibration performed using code from [1]). We calibrate over time using several frames in order to cover the space well. Figure 7 shows calibration results for this sequence.*

the planar calibration object in one portion of the scene. Even worse, such calibration objects require that every camera is able to see the planar object. This restriction can be limiting in the case of large numbers of cameras and awkward geometries.

Reprojection errors are not just a problem in theory, but also in practice. As figure 8 demonstrates, the reconstruction offered by our method is consistent with physical models of cloth. Cloth geometry reconstructed from poor calibration can be heavily strained — meaning that the distance between points is not accurately recovered. In graphics applications, strain is considered so distracting that some simulation methods explicitly contain a strain reduction step [13].

Appendix: Metric Upgrade Proof

A metric reconstruction isn't possible from two views in simple orthographic cameras without calibration of camera extrinsics or some known length or angle [8, 18]. The reconstruction ambiguity is instructive to study further. Write $\mathcal{D}$ for a view by point data matrix and $\mathcal{P}$ for a 3xpoint geometry matrix; there must be a minimum of four points. Define a **canonical two-camera matrix** to be a matrix of the form

$$\mathcal{C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ e_{11} & e_{12} & e_{13} \\ e_{21} & e_{22} & e_{23} \end{bmatrix}$$

and $\mathbf{e}_1$, $\mathbf{e}_2$ are arbitrary orthonormal 3 vectors. We move the origin to the center of gravity, absorb scale into the points, and place the first camera in canonical position to obtain $\mathcal{D} = \mathcal{C}\mathcal{P}$ where $\mathcal{C}$ is a canonical two-camera matrix. If $\mathcal{L}$ is a matrix such that $\mathcal{C}' = \mathcal{C}\mathcal{L}$ is also a canonical two-camera matrix, the

calib data (cam model)	evaluation data	reproj error
cloth: frames 1-10 (ortho)	cloth: frames 1-10	2.06 pixels
cloth: frames 1-10 (persp)	cloth: frames 1-10	0.35 pixels
cloth: frames 1-10 (persp)	**cloth: frames 11-50**	**0.68 pixels**
cloth: frames 1-10 (persp)	checkerboard	2.15 pixels
checkerboard (persp)	checkerboard	0.30 pixels
checkerboard (persp)	**cloth: frames 11-50**	**2.62 pixels**

Figure 7: *Both calibration objects (the cloth and the checkerboard) give good calibration results on the data used to calibrate and both methods generalize poorly to data in physically different locations. However, the calibration using the cloth gives significantly better results on cloth data from another part of the sequence. On unseen data, calibration from the cloth produces average errors of 0.68 pixels, while calibration from the checkerboard produces average errors of 2.62 pixels. This is because the cloth moves through the same 3D volume in one portion of the sequence, allowing a better reconstruction in the remaining portion of the sequence. The checkerboard calibration pattern has the additional disadvantage that it must be viewed by all cameras and often occupies valuable camera pixels.*

reconstructions $\mathcal{P}$ and $\mathcal{L}^{(-1)}\mathcal{P}$ are both available. Note that $\mathcal{L}$ is a matrix of the form $[[1,0,0];[0,1,0];[a,b,c]]$. A one parameter family of such $\mathcal{L}$ exists, and they are not Euclidean transformations. Now assume that we are working with patches.

Lemma: *A metric reconstruction from two simple orthographic views is available from two patch correspondences. There is a maximum of sixteen ambiguous cases, yielding no more than four camera reconstructions.*

Proof: We must first deal with scale, as the two cameras may have pixels of different sizes. Scale commutes with reconstruction, meaning that a camera with small pixels produces a larger frontal view of the texture elements. The ratio of camera scales is then found by scaling a frontal view of an element in the first camera to be the same size as a frontal view of an element in the second camera. Note that correspondences between element *instances* are not necessary to do this. Each patch consists of a point and a projected normal vector. Write the j'th point as $\mathbf{P}_j$ and the i'th view of the j'th point as $\mathbf{p}_j^i$. Write the j'th normal as $\mathbf{N}_j$ and i'th view of the j'th projected normal vector as $\mathbf{n}_j^i$. What we have referred to as the "head" of the i'th view of the j'th projected normal vector is then $\mathbf{p}_j^i + \mathbf{n}_j^i$; it is easier here to work with the vector directly. First, a metric reconstruction is available because the normal vectors are unit vectors in 3D; we can obtain the metric reconstruction by choosing the element of the one parameter family $\mathcal{L}$ that makes the first normal a unit vector. Ambiguity is more interesting. Our ambiguity in the projected normal vector is a sign ambiguity, yielding a total of sixteen ambiguous cases (two per view per patch). However, these ambiguities have an important internal structure. Write $\mathcal{D}_{(kl)}$

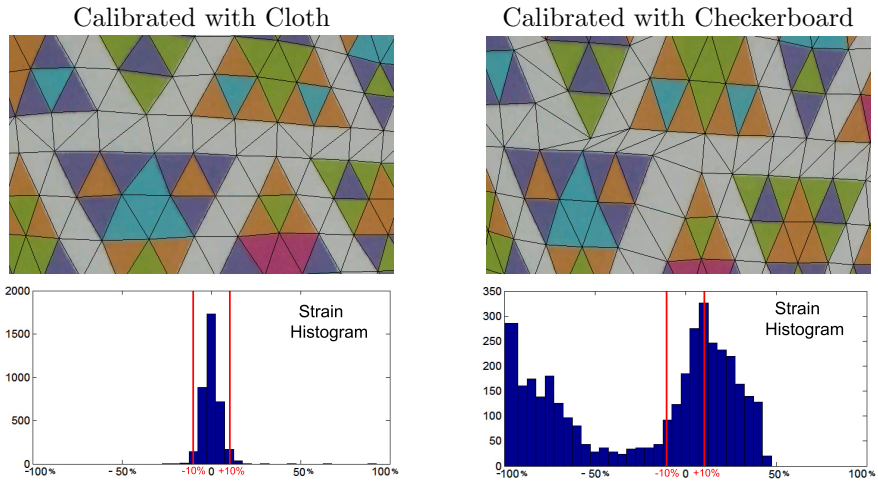


Figure 8: *We demonstrate the power of our calibration method by computing reconstructions that are consistent with the physics of cloth. On the* ***left****, a close-up view of a reconstruction using the triangle pattern in other frames to calibrate. On the* ***right****, a similar view reconstructed using the checkerboard as a calibration object. (original photos of this sequence in figure 3) Errors in the checkerboard reconstruction are two-fold: First, the relative scaling of the axes is less accurate. Second, small errors in calibration produce strain — or stretch and compression of the edges connecting neighboring vertices. Notice substantial errors on the right in the white regions surrounding each triangle.* ***Below****, histograms of the strains over the whole surface highlight the importance of good calibration. Values that deviate significantly from zero (rest length) correspond physically to large forces on the cloth. [13] notes that strains for textile materials are typically less than 10%. (Technical note: strain is* $\frac{\Delta L}{L}$*, where* L *is the rest length. To compute rest lengths, we estimate a scale factor between the model and the reconstruction. For checkerboard calibration, our estimates are dubious because we choose a scale estimate that minimizes strain. When the strains are consistent, this task is easy, but when they are inconsistent, the task becomes harder. No matter what method we use, the checkerboard calibration produces large strains)*

for

$$\begin{bmatrix} \mathbf{p}_1^1 & \mathbf{p}_2^1 & \mathbf{n}_1^1 & \mathbf{n}_2^1 \\ \mathbf{p}_1^2 & \mathbf{p}_2^2 & (-1)^k\mathbf{n}_1^2 & (-1)^l\mathbf{n}_2^2 \end{bmatrix}$$

and $\mathcal{I}^{(ij)}$ for $\text{diag}(1, 1, -1^i, -1^j)$. We then have that the ambiguous cases are $\mathcal{D}_{(kl)}\mathcal{I}^{(ij)}$ for $(i, j, k, l) \in [0, 1]^4$. Now if $\mathcal{D}_{(kl)} = \mathcal{C}_{(kl)}\mathcal{P}_{(kl)}$, then $\mathcal{D}_{(kl)}\mathcal{I}^{(ij)} = \mathcal{C}_{(kl)}\mathcal{P}_{(kl)}\mathcal{I}^{(ij)}$. This means that there are only four cases for the camera matrix. Furthermore, our ambiguities do not interfere with metric reconstruction. Note that

$$\mathcal{P}_{00}\mathcal{I}_{(kl)} = \left[\mathbf{P}_1\mathbf{P}_2(-1)^k\mathbf{N}_1(-1)^l\mathbf{N}_2\right]$$

so that for any of four cases $\mathcal{D}_{00}\mathcal{I}^{(ij)}$ we will obtain the correct camera by insisting that the third column of $\mathcal{P}$ is a unit vector. Furthermore, in these four cases the fourth column will be a unit vector, too. We do not expect this to be the case for the other twelve cases in general — though specific geometries may make it possible — so that the correct camera is generally easily identified. □

References

[1] Jean-Yves Bouguet. Camera calibration toolbox for matlab, 2005. http://www.vision.caltech.edu/bouguetj/calib_doc/.

[2] Olivier Faugeras and Quang-Tuan Luong. *Geometry of Multiple Images: The Laws That Govern the Formation of Multiple Images of a Scene and Some of the Applications*. MIT press, 2004.

[3] D.A. Forsyth. Shape from texture without boundaries. In *Proc. ECCV*, volume 3, pages 225–239, 2002.

[4] I. Guskov and L. Zhukov. Direct pattern tracking on flexible geometry. In *The 10-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2002 (WSCG 2002)*, 2002.

[5] Igor Guskov, Sergey Klibanov, and Benjamin Bryant. Trackable surfaces. In *Eurographics/SIGGRAPH Symposium on Computer Animation (2003)*, 2003.

[6] R. Hartley and A. Zisserman. *Multiple View Geometry*. Cambridge University Press, 2000.

[7] Janne Heikkila and Olli Silven. A four-step camera calibration procedure with implicit image correction. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, page 1106, Washington, DC, USA, 1997. IEEE Computer Society.

[8] T.S. Huang and C.H. Lee. Motion and structure from orthographic projections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1989.

[9] Anthony Lobay and D.A. Forsyth. Recovering shape and irradiance maps from rich dense texton fields. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2004.

[10] Marc Pollefeys, Reinhard Koch, and Luc Van Gool. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. *Int. J. Comput. Vision*, 32(1):7–25, 1999.

[11] D. Pritchard. Cloth parameters and motion capture. Master's thesis, University of British Columbia, 2003.

[12] D. Pritchard and W. Heidrich. Cloth motion capture. *Computer Graphics Forum (Eurographics 2003)*, 22(3):263–271, September 2003.

[13] Xavier Provot. Deformation constraints in a mass-spring model to describe rigid cloth behavior. In *Graphics Interface 95*, pages 147–154, 1995.

[14] V. Scholz, T. Stich, M. Keckeisen, M. Wacker, and M. Magnor. Garment motion capture using color-coded patterns. In *Proceedings of EUROGRAPHICS*, 2005.

[15] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. J. Comput. Vision*, 9(2):137–154, 1992.

[16] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *ICCV '99: Proceedings of the International Workshop on Vision Algorithms*, pages 298–372. Springer-Verlag, 2000.

[17] Roger Y. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. 1992.

[18] S. Ullman. *The Interpretation of Visual Motion*. MIT press, 1979.

[19] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(11):1330–1334, 2000.