Using thousands of images of an object

Robert Pless* and Ian Simon
Department of Computer Science,
Washington University in St. Louis

Abstract

In this paper we consider the analysis of thousands of unorganized, low resolution images of an object. With very low resolution images, standard computer vision techniques offinding corresponding points and solving for image warping parameters or 3D geometry may fail. Two recent techniques in statistical pattern recognition, locally linear embedding (LLE) and Isomap, give a mechanism for finding the structure underlying point sets for which comparisons or distances are only meaningful between nearby points. We explore these methods to simultaneously compute camera position and object pose for thousands of images using nothing but a global similarity measure between images.

1 Introduction

Surveillance applications may produce many low resolution images of a single object. With very low resolution images taken from unknown viewpoints, standard computer vision algorithms do not have a good handle to begin the image understanding process. Here, we study a problem where we have thousands of pictures of an object, parameterized by an unknown camera viewing angle and object rotation. Given enough images so that every image has similar images taken from nearby viewpoints, is it possible to extract that camera position and object pose for all the images?

One popular tool that could give an approach to this problem is called multi-dimensional scaling (MDS) [2]. Explained more in the next section, this technique is a method for creating an embedding of a point set that respects the set of all pairwise distances. Naively applying this to image data requires a meaningful measure of distance between all pairs of images. For images taken from very similar viewpoints, almost any distance metric between images will be small. For image taken from dissimilar viewpoints, almost any image distance metric is likely to be uncorrelated with the actual distance between camera viewpoints. Therefore embedding the images in a parameter space using MDS directly is not satisfactory.

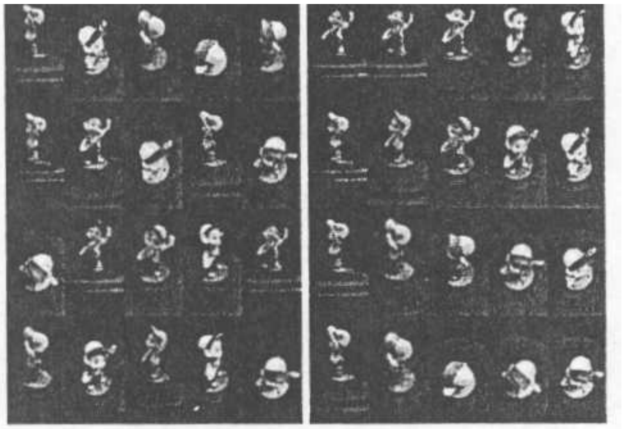


Figure 1: We consider the problem of organizing an unordered set of small images (left). Using Isomap or LLE, it is possible to automatically organize the pictures into a low dimensional parameter space, in this case a 2 dimensional space (right). The subject of this paper is to modify or extend these techniques in order to extract metric properties of camera angle and object pose.

Fortunately, two recent papers give tools to allow a MDS like solution for situations when only a local similarity measure is required[3, 6]. Each of these tools takes as input *local* relationships between input data points. Each outputs coordinates for the data points that best satisfy the given relationships. Unlike principal component analysis, these coordinates do not have to correspond to a linear subspace of the space in which the original point set lies - If the point set lies in a low dimensional *non-linear* manifold (like a spiral jelly roll), the coordinates specify point locations within that manifold.

The work presented here is an initial exploration into the use of LLE and Isomap in the analysis of image data. We explore different image distance metrics and give two mechanisms which use outside knowledge to force the parameters of the embedded point set to conform to parameters of interest in the world. In particular, an experiment in section 4 uses 1,800 images of an object and embeds the

*to whom correspondence should be addressed: plessa@cs.wustl.edu images into a space parameterized by the angle of elevation of the camera and the angle of rotation of the object.

Related to this work is [5], who does a similar dimensionality reduction by comparing a large set of images to a set of templates. Comparing the images to the templates avoids the need to compare all pairs of images, and gives a method to find a low dimensional Euclidean embedding of the image set, suitable for content based indexing. However, this still requires that a distance metric be valid for every pair of template and image, instead of a measure that need only be valid for very similar images.

2 MDS, LLE and Isomap

In this section we give a minimal overview of the mathematics behind LLE and Isomap, the subsequent sections give modifications to these procedures which to force the embedding to have physically meaningful parameters. A complete description is available in the original articles [3, 6], and longer tutorials on these methods and MDS in general give more specific implementation details [4, 2].

- Multi-Dimensional Scaling
  Input: D = $n \times n$ matrix of all squared pairwise distances
  Output: Point coordinates which best approximate pairwise distances.
  Method: Solve an eigenvalue problem with a matrix easily created from D

- Isomap
  Input: an $n \times n$ matrix pairwise distances with some (perhaps most) distances unknown.
  Output: Point coordinates such that the pairwise distances are best approximated.
  Method: Define a graph whose vertices are the set of points, and whose edges are the known pairwise distances. Compute all-pairs shortest path distances in this graph, which defines a distance between every pair of nodes. Use MDS to find point coordinates which satisfy these (now complete) distance constraints.

- LLE
  Input: A $n \times n$ weight matrix W which expresses each point as a weighted sum of other points (probably neighbors).
  Output: Point coordinates best fitting the local constraints
  Method: Solve an Eigenvalue problem to find reasonable point coordinates X such that $WX = X$.

Both Isomap and LLE output a set of point coordinates. In the subsequent section, we explore techniques to allow those point coordinates to have a physical meaning.

3 Constrained Embeddings

Extra knowledge is required to transform the embedded point set into one that expresses metric information. There are two categories of external knowledge that can be brought to bear on the embedding process. The first method is to enforce absolute knowledge of the desired parameter location for one or a small set of the points. The second is to enforce global properties of the embedding, for instance the knowledge that the data set comes from an even sampling of the desired parameterization. The form of global constraints which are appropriate is highly application dependent, and we discuss techniques used in our experiment within the experimental section.

3.1 Local Constraints

The LLE approach to embedding the point set starts with a weight matrix W which which expresses each point as a weighted sum of other points. It then seeks a set of point coordinates X which respect this weighting:

$$WX = X$$
$$(W - I)X = 0$$

Requiring that certain points must be embedded in particular locations requires the solution to a similar problem:

$$(W'-I)X'=C,$$

where X' is the remaining unknown point coordinates for which we are solving, W' is the matrix of relative constraints between these points, and C is a matrix encoded the effect of the location of the fixed points.

Alternatively, the points can be warped after the fact to force the satisfaction of a particular set of constraints. Different warping functions may be required depending upon the number of points whose position is fixed. A general linear transform is an 8 parameter transform allowing any four points to be fixed. The four embedded points, and their desired locations, define a linear system which can be solved to define the transformation for each point:

$$W, Y') = \left(\frac{ax+by+c}{gx+hy+1}, \frac{dx+ey+f}{gx+hy+1}\right)$$

4 Experiment

The visual object capture system, shown in Figure 2, captured 1,800 images of the object shown in Figures 1 and 5.

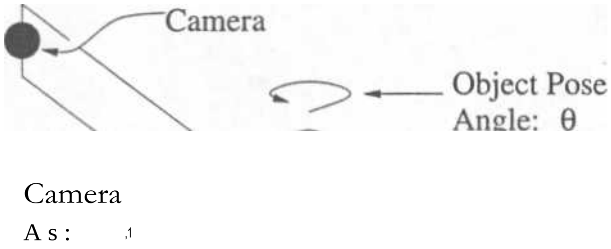


Figure 2: Object capture system, can take pictures of an object from any camera angle and object rotation. For this experiment, we took 1,800 images, sampling O and 9 every 3 degrees (data set available upon request, uncompressed pgm format).

These images evenly sample the space of object rotations (every 3 degrees, over one half a rotation) and camera viewing angle (every 3 degrees, from horizontal to vertical). The images were sub-sampled to 32 x 64 pixels, and all pairs of image distance were computed as both the sum of squares of differences of normalized pixel intensity, and the sum of squares of differences of normalized edge maps. We found that using the edge image gave qualitatively similar but slightly better and more consistent results that using distances computed from original images.

For Isomap, the local neighborhood graph of each image consisted of the 8 nearest images, all other distances were initially unknown and defined during the Isomap procedure. For LLE, each point was expressed as a weighted sum of its neighbors using the following algorithm suggested in [4]. Using all pairwise distance between a point and its eight nearest neighbors, embed these (nine) points using MDS. Then, express the central point as a weighted sum of its neighbors, and use these weights as the input constraints to LLE.

Figure 3 shows the result of the LLE embedding of all 1800 images. The four points corresponding to the (known) extremes of camera angles and object rotations are marked with small circles. In the coordinate system defined by these four points, the points should be arranged as a rectilinear grid. This metric structure underlying the point set is not exhibited by this embedding. Exploring why this fails is a subject of future work - It may be better to compute each image directly as a linear combination of neighbor rather than first computing distances, then locally embedding, then using that local structure. A qualitative structure is found, shown in Figure 1 (right); as one moves in a path through neighboring points in this embedding, there is a smooth transition between image viewpoints, but the embedding does not directly capture the parameters of the object pose.

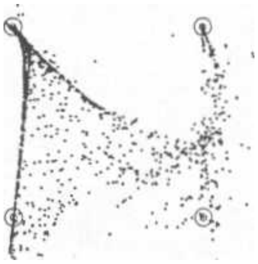


Figure 3: Locally Linear Embedding: Each image was expressed as a linear combination of nearby images. The points were embedded with the additional constraint that the four comer images lie at fixed positions at the corner of a square.

Figure 4 (top) shows results using the standard Isomap procedure. Choosing four extreme points (circled), and solving for the general linear transform which forces these four points to have specific coordinates (bottom left) allows one to define meaningful axes to the embedded space. Finally, the a-priori knowledge that the parameter space was evenly sampled gives a global constraint on the embedded point set. Using a variant of the thin-plate spline warping technique [1] to enforce that the density of points in every region of the embedded space is approximately constant gives the final embedding (bottom right). Evenly sampling this space and choosing the closest image to the sample points gives a graphical depiction of this embedding (Figure 5).

Finally, since this data set was taken in a laboratory setting, the actual pose coordinates are known for each image. Over all 1800 images, the mean error in the embedded 0, O coordinates was: 6.98 °, 2.97°. This numbers should not be compared directly to other pose estimation algorithms, and are extremely good, given that they come from the analysis of 32 x 64 pixel images of an unknown object.

5 Conclusions and Future Work

The tools of Isomap and LLE give a fascinating new approach to finding statistical relationships in large point sets. Applying these techniques to vision applications holds to promise of solving problems for which there is currently no good approach. However, the domain of sets of images is a complicated one. Distances between images (by any metric) do not have a consistent relationship to changes in camera position or object pose. Explicitly

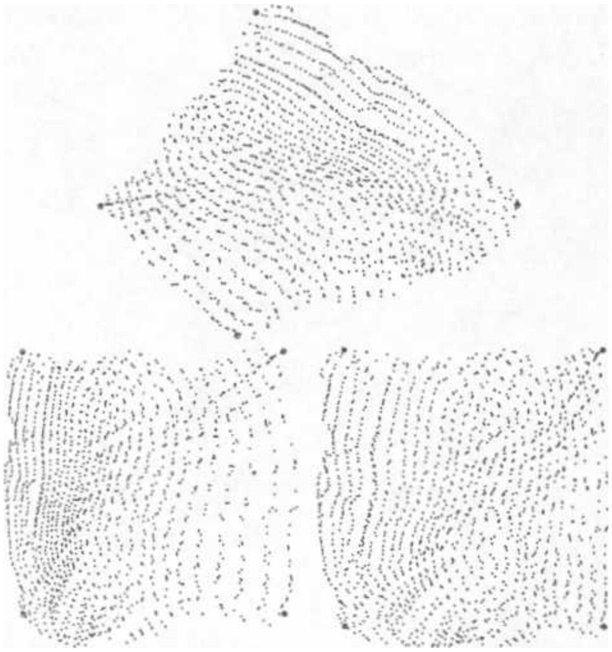


Figure 4: At the top is the initial Isomap solution embedding the positions of all 1,800 images. In the bottom left, this embedding is warped by a general linear transform so that the four circles points lie in fixed points in the final embedding. In the bottom right image, a variant of thin plate spline enforces the constraint that the parameter space was evenly sampled. The horizontal axis of this plot corresponds to the camera angle 0, and the vertical axis is the object rotation angle 0.

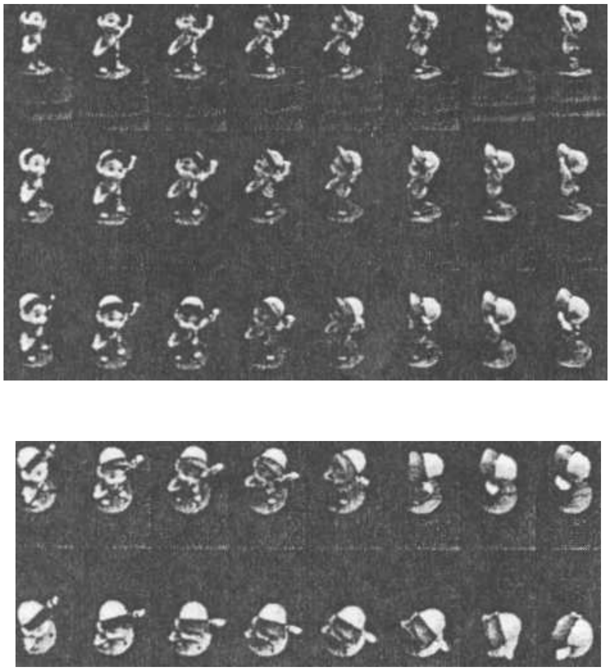
imposing external knowledge during the embedding process or warping the embedded coordinates allows approximate pose estimation.

Future work should explore better mechanisms to put external knowledge into the LLE embedding. Also, different image metrics should be explored; for particular properties that one wants to extract, direct comparison of image pixel intensity may be inherently less suitable than a distance measure based upon optic flow, feature points, or hierarchical models. Finally, other classical computer vision questions such as structure from motion and segmentation should be explored in the context of simultaneously using thousands of uncalibrated images.

References

[11 F. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11:567-585, 1989.*

t h"



5: Images from an even sampling of the final embedded space shown in Figure 4, (bottom right). The mean embedding error over all *1800* images is only 6.98° for the rotation, and 2.97° for the camera elevation.

[2] I Borg and P J F Groenen. *Modern multidimensional scaling theory and applications.* Springer, New York, *1997.*

[3] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science, 290:2323-2326,200.*

[4] Sam T Roweis and Lawrence K Saul. An introduction to locally linear embedding: http://www.gatsby.ucl.ac.uk/~roweis/lle/papers/Ileintro.pdf , *2001.*

[5] Haim Schweitzer. Template matching approach to content based image indexing by low dimensional eu~ clidean embedding. In *Proc. International Conference on Computer Vision,* pages *566-569, 2001.*

[6] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science, 290:2319-2322, 2000.*