MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Matching Images of Urban Scenes

Paul Beardsley

TR2003-09 February 2003

Abstract

This report describes a preliminary study on the problem of matching a live camera image of an urban scene to a small set of images, retrieved for the camera's current geographical location from a database, to find the most similar view. The approach is feature-based, first detecting point features in the images, then identifying potential point matches based on feature similarity, and finally applying the fundamental matrix constraint to retain good matches and remove poor matches. Results of image matching are shown for scenes under the same and varying illumination.

Created February 2003.

Matching Images of Urban Scenes
Technical Report TR2003/09

Paul Beardsley

Mitsubishi Electric Research Laboratories,
201 Broadway, Cambridge MA 02139, USA,
pab@merl.com,

Abstract. This report describes a preliminary study on the problem of matching a live camera image of an urban scene to a small set of images, retrieved for the camera's current geographical location from a database, to find the most similar view. The approach is feature-based, first detecting point features in the images, then identifying potential point matches based on feature similarity, and finally applying the fundamental matrix constraint to retain good matches and remove poor matches. Results of image matching are shown for scenes under the same and varying illumination.

1 Introduction

The application of this work is a cellphone with camera and display, with the capability to automatically recognise an urban scene in the camera's field of view in order to present an augmented view. In the envisaged system, the cellphone uses GPS to obtain its approximate location. A database is then accessed to obtain pre-recorded images of the surrounding buildings. The system compares the live camera image with the database images, and having found the most similar view, the cellphone can display the live image of the scene overlaid with any augmentation data available for that view.

Thus the system effectively uses a coarse-to-fine approach to identify a scene, with GPS providing approximate location and vision doing the fine-tuning. Current GPS has an accuracy of about 10m in wide-open environments, but this degrades or breaks completely in urban environments due to obstruction and reflection. Differential GPS is a less widely-available service but is more accurate, say 1m in wide-open environments but again degrading in urban environments. As for orientation of the device, an approximate estimate can be obtained relatively cheaply using compass and tilt-sensor. Again urban environments can cause problems due to, say, stray magnetic fields from the metal hulls of buildings.

For the purpose of this study, the assumption is that there is GPS but no orientation information available. The GPS data is the basis for retrieving from a database all images for the 360° surroundings at a location, and vision refines this by matching the live camera image to one image from the 360° set. The prototype

vision system is based on feature matching using the geometric constraint of the fundamental matrix. A robust real-world system would certainly require more than this, and we outline the most immediate extensions in the future work section.

That said, the prototype system does demonstrate successful matching on the sample data provided for the study, six datasets of urban environments in Japan. The sample data shows a variety of urban scenes. But in fact there is no change of illumination between the source images and database images, so it does not reflect what would happen in a real application. We added one dataset where illumination does change between the source and target images, using images of city locations which were downloaded from the web. These images are taken with different cameras as well as under different illuminations, and there are long-term changes such as vegetation change in the surroundings. This dataset demonstrates some successful matching while also indicating the limits of the current system.

2 Overview

The goal is to match a source image of an urban scene against a set of target images to find the most similar view. For initial preparation of the data, images are resampled if necessary to be approximately the same size. This ensures that physical features will have similar appearance in different images, at the pixel level.

The source image is matched against each target image in turn as described in Section 3. The results of the pairwise matching are used to rank the target images according to similarity to the source image, as described in Section 4.

3 Pairwise Image Matching

There are three stages to the algorithm for matching between a source image and a given target image -

- corner detection using the Harris corner detector [3],
- computing zero or more hypothesised matches for each corner, based on gradient similarity around the corners,
- applying the fundamental matrix constraint to retain good matches and remove poor matches.

3.1 Pairwise Image Matching - Computing Hypothesised Matches

Each corner $c1$ in the source image is processed in turn. For $c1$ having coordinates (x, y), the search area for corresponding corners in image 2 is a circle centred on (x, y) with a radius r which reflects the maximum possible disparity between the images. A match score is computed between $c1$ and each corner $c2$ in the search area, as described in Section 3.2. If the match score is worse than a threshold

δ, the match is rejected, else the match is stored. The acceptance threshold is made very lax since our goal is to ensure that the true match for $c1$ is very likely amongst the collected set of hypotheses, even if this means that each $c1$ has many hypothesised matches. Pruning the hypothesised matches will take place later during computation of the fundamental matrix.

3.2 Pairwise Image Matching - Scoring an Individual Match

This section defines the match score for a pair of corners c_1 and c_2 in images 1 and 2 respectively. The scoring utilises gradient information rather than color, to obtain invariance to illumination change. For an RGB image, the gradient is computed in a single channel.

The process computes a match score between a square patch of pixels centred on $c1$, and a corresponding patch at $c2$ but offset by (o_x, o_y). The offset values range from zero to a maximum search offset. The best score over all offsets is retained as the final match score. For this system, the patch size was 5x5 pixels, and the maximum search offset around $c2$ was 1 pixel.

The match score s at a particular offset is

$$s = \sum_p s_p \tag{1}$$

where p are the pixels in the patch. The score s_p for each individual pixel location in the patch is zero if there is zero gradient at either pixel 1 or pixel 2 for that location, else

$$s_p = g_{v1}.g_{v2} * r * g_{min} \tag{2}$$

where

g_{v1} is the pixel 1 gradient unit-vector
g_{v2} is the pixel 2 gradient unit-vector
r = min-value $(g_{m1}/g_{m2}, g_{m2}/g_{m1})$
g_{min} = min-value (g_{m1}, g_{m2})

and

g_{m1} is the pixel 1 gradient magnitude
g_{m2} is the pixel 2 gradient magnitude

In this matching procedure, we seek to reward corresponding pixels with an aligned gradient direction and with similar magnitudes, hence the dot product and the r term in equation (2). We also seek to downgrade the score if one or both gradient magnitudes is small, since small gradients are unreliable. The use of g_{min} rather than $g_{m1} * g_{m2}$ reflects the fact that only one of two corresponding pixels need have a small gradient magnitude to render its contribution unreliable. When both gradients are equal, $s_p = g_m$ so pixels with higher gradients make a higher contribution.

3.3 Pairwise Image Matching - Computing the Fundamental Matrix

The fundamental matrix F is a geometric constraint between corresponding points in a pair of images. Given F, and a point in image 1, its corresponding point is constrained to lie on a line in image 2.

Given a set of putative matches between corners in image 1 and image 2, and no prior knowledge of F, a typical strategy is to identify an F which agrees with as many of the matches as possible, and to discard all inconsistent matches [4]. For our case, each corner $c1$ in image 1 has zero or more hypothesised matches in image 2, and the goal is to identify an F such that as many $c1$ as possible have at least one hypothesised match which is consistent. Inconsistent hypothesised matches are culled. After this culling there may still be some cases where some corner $c1$ has more than one hypothesised match, and we use the match score to identify and retain the best match.

The computation of the fundamental matrix is based on RANSAC,[2], which involves taking repeated subsamples of matches from the full set, computing an F, and checking its agreement with the full set, to identify good estimates of F. There is a modification to the sampling in the following way.

For each corner $c1$, the hypothesised matches are sorted according to score. Samples are initially taken only only from the top-score hypothesis at each $c1$. Once some solutions for the fundamental matrix start to be generated, then the hypothesised matches are also sorted according to geometric agreement with the current estimate of the fundamental matrix. Now, alternate samples during the sampling process are taken from the top-scoring hypotheses and the current most geometrically consistent hypotheses.

Even images of different scenes can give apparent success in the computation of F due to the occurrence of matches which accidentally agree with some F. However only images of the same scene are likely to have the property that they successfully generate an F *and* the scores of the accepted matches are good. This is the subject of the next section.

4 Ranking the Target Images

The previous sections described the matching process between the source image and each of the target images. The goal now is to rank the target images in order of similarity to the source. For this study, we have made the assumption that the source image does match at least one of the target images. Thus the goal is find the best-matching target image, rather than the more difficult task of making a binary match or not-match decision for each target.

There are two indicators of what constitutes a good match between the source image and a target image. The first is the number of matched points, and the second is the associated match scores. The ranking is performed in the following way. First the maximum number of matches k for any source-target image pair is identified. Any source-target image pairs with less than $k/2$ matches are eliminated from further consideration. For all remaining source-target image

pairs, the matches are sorted according to match score, and the score at the $k/2$ element is used to rank the targets.

5 Results

The sample data consists of six datasets of urban environments in Japan, with a total of 15 source images and 42 target images. Three source images were eliminated from the test set because there was high zoom (5x) between the source and target, something not handled by the current algorithm. Eleven of the remaining twelve source images were successfully matched, with one failure due to the presence of a target which was similar to the source without actually being the true match. The system matches images at about 1Hz, typical image size 160x120 pixels, although this is with miscellaneous GUI output occurring, and no attempt has been made to optimise the code.

5.1 Matching Results

Figure 1 shows a typical source image and a set of target images.

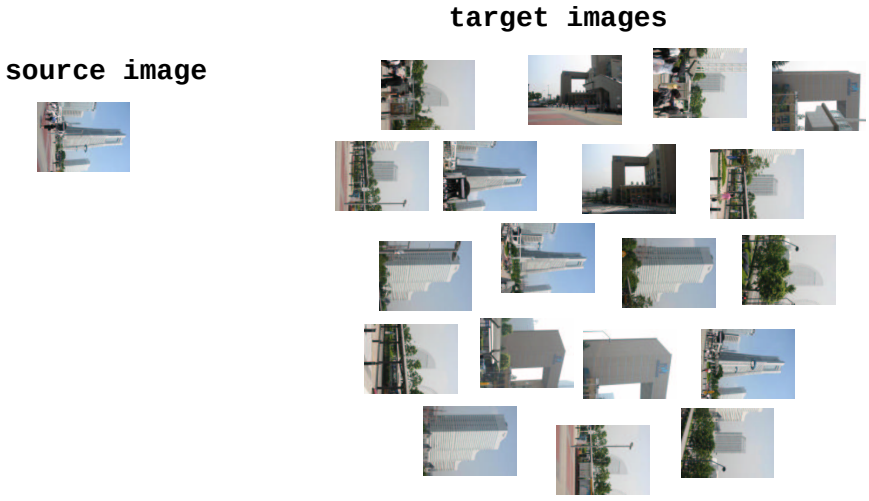

Fig. 1. A typical source image and a set of target images.

Figure 2 shows the matching results for all source images in the sample data.

Figure 3 shows, for the source image in Figure 1 taken pair-wise with each target image, the distribution of corner match scores for accepted matches. In fact there are results for only 14 source-target pairs, since five pairs were automatically eliminated as unacceptable during the matching stage. Figure 3 shows

Fig. 2. Each pair of images shows a source image at left and the highest rank target image at right. All pairs above the horizontal line are correct matches. The pair below the horizontal line is incorrect.

no quantum leap in the scores for source-target pairs where the source and target show the same scene, and source-target pairs with different views. This is a reflection of the limitations of the current system, since a strong division between the results for valid image matches and incorrect image matches would be evidence of robustness, but is not apparent.

Figure 4 shows a dataset in which the images were taken with different cameras under various illuminations, and Figure 5 shows the matching results.

5.2 Developer Information

The prototype application, called UrbanMatch, is based on the Diamond3D vision library. An introduction to the library is given in [1]. The high-level control of the application is in d3d-subsystem / corner-match-stereo2. The code for matching corners is in d3d-corner / VlCornerMatch. The code for computing the fundamental matrix using the corner matches is in d3d-corner-geom / VlCornerComputeFMatrix.

6 Future Work

The current system is based on feature matching using the fundamental matrix. There is another well-known geometric constraint which applies when a scene contains planes, as urban scenes typically do, and that is the planar homography. Unlike the fundamental matrix, a homography is not a global property of

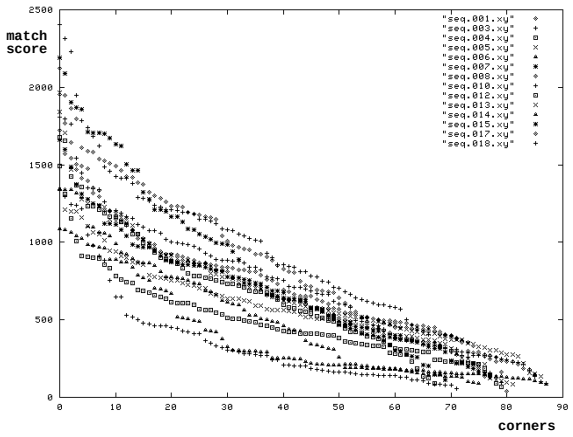

Fig. 3. Distribution of corner match scores for accepted matches, for the source image in Figure 1 taken pairwise with 14 of the target images. The graph is too cluttered to discern individual tracks but the main point is that there is not a large disparity between scores for source-target pairs where the source and target show the same scene, and source-target pairs with different views.

all rigid parts of the scene but applies only to the features on a given plane. A key advantage is that it applies both to points and lines, so it provides an invaluable aid to line matching. There are various ways to utilise homographies in conjunction with the fundamental matrix as part of an integrated matching scheme, and this would be the best next step for the work.

Another important step is to get a better understanding of how features change appearance under changing illumination. The current algorithm attempts to achieve some invariance to illumination change by doing corner matching based on gradients (Section 3.2). But this needs a fuller investigation, examining how features in a scene appear under, say, five-six different illumination conditions, and determining what is invariant.

In the longer-term, a likely step would be to add area-based matching. This would make explicit the surfaces in the scene, as well as providing more reliable verification for the matching.

7 Conclusion

This work is a preliminary study of the problem of matching a live image of an urban scene with database images of all the surrounding buildings at that location. The prototype system demonstrates the viability of the feature matching approach, even under illumination change or with changes in the physical scene such as presence and disappearance of vehicles, vegetation change etc. The work represents only a first step toward a true working system, and we outlined the primary target for any next stage of the work - utilising homographies to

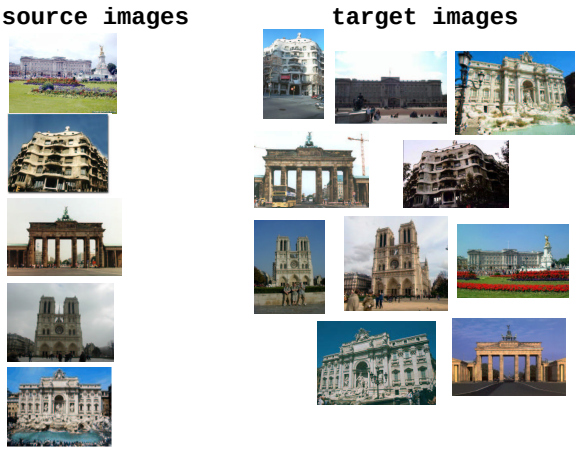

Fig. 4. There are five source images, each of which is matched against the set of 10 target images. These scenes show Buckingham Palace, Casa Mila, the Brandenburg Gate, Notre Dame, and the Trevi Fountain.

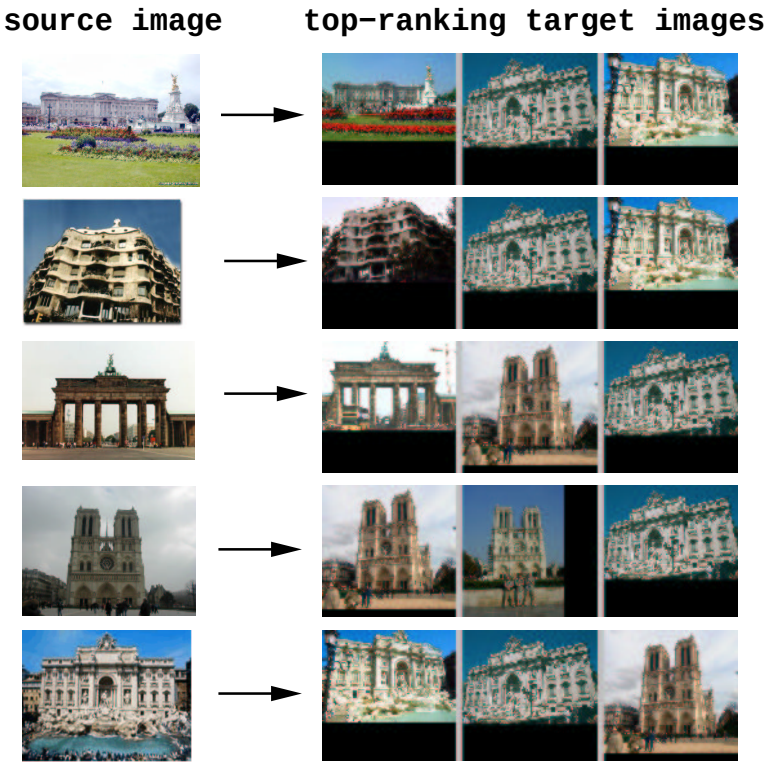

Fig. 5. Each row shows, at left a source image, and at right the top three ranking target images. The top-rank match is correct in all cases. The second-rank match is correct only in the bottom two cases.

match planes in the scene - as well as indicating a roadmap for longer-term development.

References

1. P.A. Beardsley. Introduction to the Diamond3D Vision Library. TR 2003/10, MERL, 2003.
2. M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Commun. Assoc. Comp. Mach.*, vol. 24:381–95, 1981.
3. C.G. Harris and M. Stephens. A combined corner and edge detector. In *Fourth Alvey Vision Conference*, pages 147–151, 1988.
4. R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.