Toward video to geospatial reference image indexing

Vitaly Zholudev

Richard P. Wildes

Technical Report CS-2006-03

May 15, 2006

Department of Computer Science

4700 Keele Street North York, Ontario M3J 1P3 Canada

Toward video to geospatial reference image indexing

Vitaly Zholudev
Richard P. Wildes
Department of Computer Science and Engineering
and the Centre for Vision Research
York University
Toronto, Ontario M3J 1P3
Canada

Abstract

In this report, we are concerned with registration of data in geospatial databases, especially with registering images taken by different sensors and from different viewpoints of the same scene. Extant approaches break down as viewpoint and/or sensor vary beyond relatively small changes. Of particular interest in the current work is the development of techniques that allow an aerial video to index corresponding spatial location within a larger reference orthoimage, without detailed a priori knowledge of the relative acquisition scenarios (e.g., lacking telemetry). Such an approach can extend significantly the operational range of video to reference registration, as extant techniques make strong assumptions about the availability of good initialization.

We present a uniform approach to representing video and reference imagery and for quantifying the goodness of match between two image samples, one captured from each type of source imagery, that have been brought under our representation. The approach combines image appearance, characterized in terms of texture defined regions, and image geometry, characterized in terms of relationships between textured regions. By construction, the matching methods are robust to a range of photometric and geometric distortions between image sources, including changes in greylevel contrast and affine geometric transformations. In application, the developed approach can serve to structure a reference image database that can be indexed directly via similarly represented video. Empirical investigations with real and synthetic data suggest the promise of the approach.

1 Introduction

1.1 Motivation

In this report, we are concerned with registration of data in geospatial databases, especially with registering images taken by different sensors and from different viewpoints of the same scene. This research has many applications in creating and updating maps, surveys and other geospatial data sources. Additionally, this work has applications in medical radiology, including registration of CT or X-ray images taken at different time instances and from different viewing angles as well as multimodal registration (e.g., CT to magnetic resonance images).

While considerable work has been performed in image registration [4]; extant approaches break down as viewpoint and/or sensor vary beyond relatively small changes. As one particular tack on extending the range of images that can be registered, we are interested in the development of image representations that accentuate geometric and textural commonalities across different images of the same scene. Success in these investigations will enhance our understanding of multi-source image registration and extend its applicability in geospatial and medical imaging.

To be more specific, assume we have a map of North America that is composed of a large number of satellite images, which were acquired some time ago. As well, we have a recent flyover video that was acquired from an airborne platform. The goal is to update the large map using the new imagery. If the camera were perfectly geolocated with respect to the map (both position and orientation provided via on board sensing, e.g., via GPS, INS, etc.), then it would be a straightforward matter of computer graphics to project the video onto the map. In reality, telemetry accuracy and precision will be limited. Therefore, some image-based registration is needed for final alignment of video to map coordinates [28].

A significant challenge arises when a priori knowledge regarding relative position and orientation of the video platform with respect to the reference (e.g., as provided by telemetry) does not provide sufficient accuracy to initialize image-based registration. For example, extant image-based registration technology has only been demonstrated to support video to georeference image alignment when initialized to within several hundred pixels of the correct result [28]. Significantly, due to errors, drop outs and otherwise limited availability of telemetry, ineffective a priori knowledge of relative video to reference image alignment is a real-world problem.

In formulating a response to the challenge of limited a priori knowledge of relative video to reference alignment, it is important to realize that simple search strategies are not applicable; the reference imagery is too large to support effective search. As an alternative, we propose to investigate an approach that allows video-based image descriptors to index directly into a reference image database. Successful indexing into the database will imply that (at least approximate) position of the video in the reference has been recovered; this information can then serve to initialize extant technology for video georegistration.

Development of an image-based scheme for indexing position in a reference image must respond to the challenges that arise as the imagery sources (video and geospatial reference) differ in appearance due to geometric and photometric distortions. Viewpoint changes alter the geometric relations between common features across the image sources. Diurnal and seasonal variation as well as sensor sensitivity differences (e.g., visible vs. infrared) alter image photometry. As specific examples: Aspect ratio of viewed objects will appear differently in obliquely captured video in comparison to orthorectified reference imagery; the imaged contrast of mountain ridges can fully reverse in morning vs. afternoon acquisitions. In our work, it is shown that judicious combination of (i) regional descriptors that abstract from details of photometry to concentrate on underlying pattern structure (e.g., image texture) and (ii) geometric relationships between nearby regions that are invariant to viewpoint change, respond to the noted challenges in matching video to reference. In this work, we concentrate on the design and implementation of a uniform approach to representing video and reference imagery and for quantifying the goodness of match between two representations, one captured from each type of source imagery. For video, we demonstrate with capture from an aircraft with moderate obliquity (e.g., between nadir and 45 degrees off the horizon) and ground sampling distance (e.g., between 0.5-2.0 meters/pixel). For reference imagery, we demonstrate with orthorectified imagery at 1 meter/pixel. Both imagery sources are greylevel visible. This work lays the groundwork for follow-on research to develop a complete approach for video to reference indexing.

1.2 Related Research

Basic background in image registration is covered in the review paper [4]. Previous research in video georegistration (provided an initial index of video position within the reference) is described in, e.g., [28]. Related research in the organization of spatial databases is reviewed in [12]. Of most interest in the context of the current report is previous research that is concerned with image indexing. In this light, previous research can be divided into approaches that consider global statistics of image appearance, approaches that consider geometric relationships between local features that have been detected in the image and approaches that combine both appearance and geometry.

Purely appearance-based approaches endeavour to characterize an image (or objects that are depicted within an image) in terms of global statistics of simple image properties. As a specific example, in [26] the authors propose an efficient method for indexing based on colour histograms (coarsely binned color distributions accumulated across an image or region of interest). The authors proposed that color and its distribution was sufficient for complex recognition tasks. Such an approach has a number of desirable properties, such as coarse filtering that allows us to distinguish easily between red and blue objects. Further, it is not sensitive to localization errors: If an object has a particular distribution of colors, proportions of red and white representative of the Canadian flag, then it is not important to precisely

localize the maple leaf within the overall image relative to the flag boundaries to recognize that it is an instance of the class of interest. However, such extreme invariance to geometric configuration within in an image also is a weakness: Objects that differ based primarily on the spatial distribution of components will be confused. Continuing with the flag example, the Danish flag would be confused with the Canadian flag because they have roughly the same proportions of colours, even though the spatial configuration of coloured regions differ markedly. Further, a change in the hue of an imaged Canadian flag would prevent it from being recognized, as distinguishing geometry is ignored.

Another image appearance based property that can be used for image indexing is image texture [10, 14, 30]. Still, regional descriptions of image texture, without any consideration of geometric layout within the image is plagued by the same concerns as purely colour-based descriptors.

Purely geometry-based approaches to image indexing endeavour to characterize an image (or objects that are depicted within an image) in terms of geometric relationships between the location of features that are extracted from an image, without any explicit regard for feature appearance (beyond their image position). A wide variety of such approaches have been developed (see, e.g., [13] for a detailed review). In the following, a few representative approaches are described.

Affine invariant recognition schemes have been proposed, where one point is expressed as a linear combination of another three [19]. In particular, the other three points produced a basis set of 2 vectors, and their combination was used to express the fourth point, yielding two numbers that are affine invariant. Indexing schemes also have been proposed that are based on accumulations of parameters that define local linear characterizations of features (e.g., orientation as well as position) [15, 3]. In addition to point and line features, features defined in terms of corresponding regions have been used to solve model pose estimation in terms of half-plane constraints [2]. More complex features, e.g., planes, spheres and cylinders also have been used in model indexing schemes [9]. In practice, a major limitation of purely geometry-based indexing schemes arises due to the inevitability of localization errors in feature extraction: Such errors lead to corresponding indexing errors. Indeed, localization errors often are amplified in indexing due to intervening nonlinear operations (e.g., construction of the index itself from feature coordinates).

Both purely appearance and purely geometry-based approaches to model recognition and indexing are error prone. Interestingly, the individual strengths and weaknesses of these approaches can be complimentary. Appearance-based approaches sacrifice too much by completely ignoring geometric relationships within regions of analysis; however, the appearance descriptors can be powerful in providing a coarse indication of image composition. Feature-based approaches are overly sensitive to localization errors and ignore the discriminatory power that might be offered by appearance-based descriptors. With these observations as motivation, several researchers have sought to combine both appearance and geometry in their indexing and recognition approaches. In [21] the authors claim that the overall organization of a scenes

parts strongly influences its interpretation. For example, it makes sense to represent the fact that for a typical beach scene, the sky (blue) is always above the ocean (green-blue) and the ocean is above the sand (yellow). When looking for a beach scene among other scenes, looking for green above yellow and for blue above green would do well. Thus, weak geometric relationships (essentially, ordinal) are augmented with appearance (colour) to provide an approach to scene indexing. Within the context of graph-based methods, models and image data have been represented as graphs with nodes capturing appearances while edges capture geometric relationships and recognition embodied via graph matching [6]. In one of the more successful recent approaches to object recognition, local descriptors of features (in terms of characteristic scale and orientation) are grouped into clusters to recognize objects and estimate their pose [22]. Appearance and geometry also have been combined implicitly through iconic templates, practical implementations of such approaches have appealed to hierarchical, coarse-to-fine search strategies [5, 23]. Another related approach makes use of a video sequences mosaic to index individual frames [16]; this work is different from ours as both the reference mosaic and video are derived from (exactly) the same source.

Finally, it is interesting to note that approaches have been developed for indexing a newly acquired video into a database of previously acquired videos, e.g., [29]. In such cases, it is possible to make use of motion information across the videos to build useful indices. This type of approach is not directly applicable to the challenge at hand in the current investigation, because the reference imagery in the geospatial database does not contain a dynamic component that can yield motion information.

In the light of previous research, the approach that is presented in the current report makes use of combined information derived from image appearance and geometry to represent and match aerial video to a reference image. As with other recent approaches that have opted for such combination, the motivation comes from the fact that geometry and appearance are complementary in nature. To capture image appearance, standard methods for characterizing image texture in terms of its multiscale orientation structure are employed [10]. To capture geometric relations between texture defined patches, previous work in affine invariant recognition are employed [19]. To exploit the complementarity, the geometric descriptors are used as an initial filter, which is then followed by appearance to complete matching between video and reference. It appears that the developed approach is novel in its particular method for combining appearance with geometry and for its application in video to reference matching.

1.3 Outline

This report is divided into five major sections. This first section has served to motivate the problem of video to reference indexing and to place it in the context of previous research. The second section introduces a preliminary approach and corresponding experimental evaluation. Section 3 presents various refinements to the preliminary approach. Section 4 presents a set of experiments that provides further empirical support for the developed approach, including

the refinements. Section 5 presents an overall discussion. Finally, a series of appendices sketch related alternative approaches that were considered during the course of our investigations.

2 Preliminary approach

2.1 Overview

In this section we describe a preliminary approach to representing both video and reference imagery as well as to quantifying the goodness of match between two recovered representations. The approach makes use of both regional appearance descriptors and geometric relationships between regions. Image appearance is characterized in terms of texture descriptors. Texture is employed as it is manifest in any imagery source (e.g., unlike colour, which depends on capture of multiple spectral bands). In this section we do not address issues of scale and rotational invariance in the appearance descriptors; these refinements are addressed in Section 3. In this section we also introduce a simple one-dimensional approach to capturing image geometry (e.g., as might be appropriate for imagery captured from a plane on a straight-line flight path). In Section 3, we generalize our geometric representation to encompass two-dimensions.

2.2 Appearance representation

Image appearance is captured in terms of texture. Following current standard methods in computer vision for characterizing image texture, the texture representation used is defined in terms of statistics of energy at different scales and orientations [10]. In particular, given a region of interest, the image data is filtered with a set of quadrature-pair bandpass filters, tuned for multiple scales and orientations. The derived images are pointwise rectified (squared in our case) and combined in quadrature to eliminate phase variation and produce a measure of energy at particular scales and orientations. In particular, we have used the second derivative of Gaussian filters $G^{(2)}_{\theta,\sigma}$ at orientation θ and scale σ and their Hilbert transforms, $H^{(2)}_{\theta,\sigma}$ [11], to produce a local measure of energy, $E_{\theta,\sigma}(x,y)$ within a scale σ and orientation θ band, according to

$$E_{\theta,\sigma}(x,y) = (G^{(2)}_{\theta,\sigma} * I(x,y))^2 + (H^{(2)}_{\theta,\sigma} * I(x,y))^2 \tag{1}$$

with $*$ symbolizing convolution, $I(x,y)$ the images to be filtered at scale σ and orientation θ, and (x,y) image coordinates. Finally, each image is pointwise normalized through division by the sum of all the filtered image values at the same point:

$$\hat{E}_{\theta,\sigma}(x,y) = \frac{E_{\theta,\sigma}(x,y)}{\sum_{\sigma' \in \{scales\}} \sum_{\theta' \in \{orientations\}} E_{\theta',\sigma'}(x,y)} \tag{2}$$

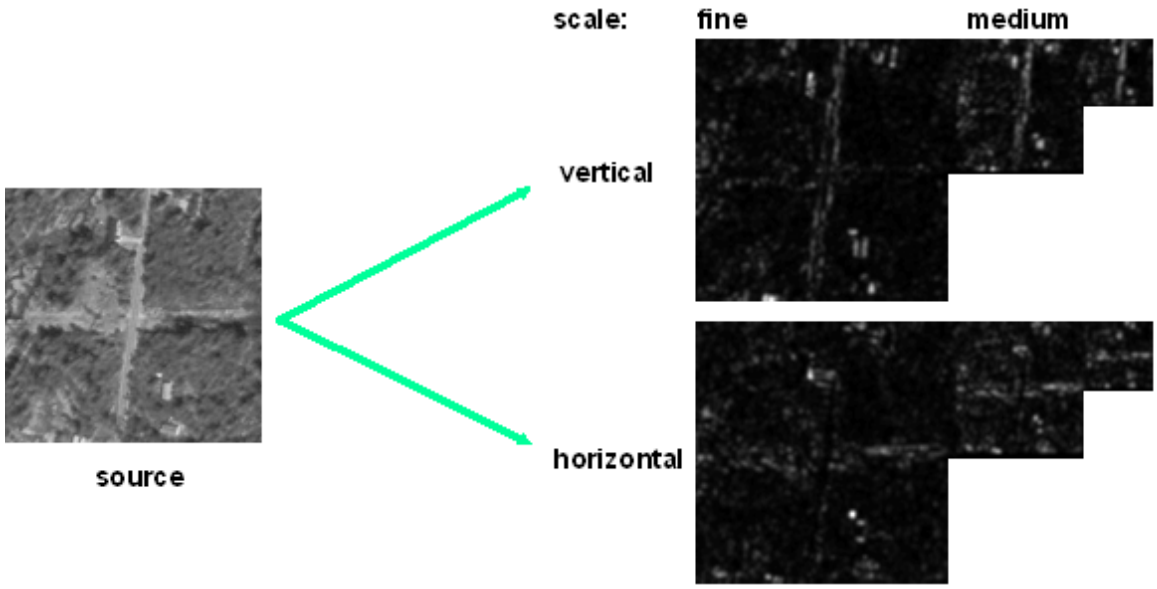

Figure 1: Multiscale Oriented Energy Image Representation. The left panel shows a source image. The right panels show derived energy images. The upper and lower right panels show local energy at vertical and horizontal orientations, respectively. Note, e.g., how vertical and horizontal structures in the source imagery are differentially highlighted in the corresponding energy images. Successively smaller energy images capture information at coarser and coarser spatial scales. In practice additional orientations and scales are employed, see text.

so that $\hat{E}_{\theta,\sigma}(x,y)$ is the pointwise normalised energy image. This normalization procedure provides for invariance to local image contrast, as local energy in each band is given as a percentage of total energy across all bands. The results yield a set of normalized energy images that parse the original image data according to scale and orientation with robustness to particulars of local image photometry, see Figure 1. In our current implementation, we decompose an image region into four scales (taken as levels 0-3 in a Gaussian pyramid [21]) and four orientations (taken as horizontal, vertical and two diagonals). Significantly, four orientations span the space of orientation for the order filter that we are using; four scales have been selected based on their discriminatory power in preliminary experiments.

Given the multiscale, multiorientation energy images that have been recovered for a region of interest, we accumulate information across the entire region by computing the mean

response,$\bar{E}_{\theta,\sigma}(x,y)$, within each band according to

$$\forall \theta, \sigma : \bar{E}(\theta,\sigma) = \frac{\sum_i \sum_j \hat{E}_{\theta,\sigma}(i,j)}{\sum_i \sum_j 1} \tag{3}$$

where i and j range over the dimensions of the texture patch. This operation yields a 16 dimensional vector (4 scales x 4 orientations) that serves to represent the appearance of the region. This same representation is used to characterize image appearance of selected regions in a video of interest as well as reference orthoimages.

Two additional points are of note. First, prior to extracting texture descriptors from a video, the individual frames are composited into a mosaic [16] to allow for spatial support to be extended beyond that of any single frame. Second, in our current investigations, we hand specify regions of interest for subsequent fully automatic analysis. Here we rely on the availability of previous research that has addressed issues in automatic (e.g., texture) segmentation that we will exploit in the future [10]. Importantly, even though automatic segmentation techniques are far from perfect, we speculate that our combined appearance/geometry approach to representation and matching will be tolerant to imperfect segmentation as it is motivated by a concern for robustness to localization and detection errors.

2.3 Geometry representation

In our preliminary approach we use a one dimensional geometry representation that we refer to as the similarity ratio, which is defined in Figure 2. Given three regions of interest, we extract the least squares fit to a line with respect to the region centroids and take the perpendicular projections of the centroids on the resulting line. The similarity ratio gives the proportion of distance the middle region projects along the segment bounded by the outer regions. The maximum distance between the central patch and the bounding patches is taken to ensure invariance to direction of traversal. The similarity ratio is invariant to shift, scale and orientation in the plane and thus can capture differences between the video and reference arising from capture at different spatial scales and (in plane) rotations.

Given three (texture defined) image regions, the similarity ratio is used to capture the relative geometric relationships between the regions. This representation is employed to capture geometric information both in videos and reference images of interest. As in the extraction of texture descriptors, video frames are mosaiced into a single large format image prior to calculation of geometric descriptors; this process brings all regions of interest into a common coordinate frame for making the relevant measurements. As with other geometric descriptors, the similarity ratio is expected to be sensitive to localization errors in its defining coordinates (here, the region centroids); however, the goal in the present work is to buttress this weakness by providing greater discriminatory power via a coupling to texture-based appearance descriptors.

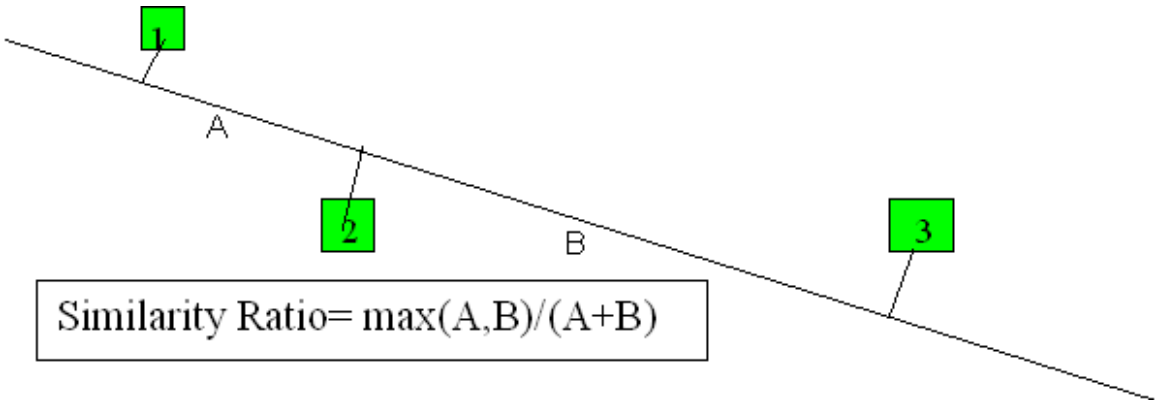

Figure 2: The Similarity Ratio. Given three patches and their centroids in the plane, the similarity ratio is defined with respect to the centroid projections on a line and comes as a ratio of distances that is invariant to the similarity group of actions.

2.4 Combining appearance and geometry

In the current approach, geometry is used as prefilter for appearance as video and reference image instances are compared. In particular, a preliminary comparison is made based purely on coarsely quantized geometry. Subsequently, only sets of regions whose relative geometry is consistent between video and reference is considered for final appearance-based matching.

The dynamic range of the similarity ratio is [0.5, 1.0], i.e., the values that the similarity ratio can take range from 0.5 to 1. It is close to 0.5 when patch 2 is half way between patch 1 and 3 and it is closer to 1 otherwise. In order to use the geometry as a prefilter for appearance we need to be able to prune entries that have different similarity ratios. At the same time, we do not want to rely heavily on high precision in the geometric descriptor. This is achieved with a simple binning technique. We divide the similarity ratio dynamic range, [0.5, 1.0], into N overlapping bins (N=7 with 50% overlap in the reported experiments). We characterize a reference image database entry in terms of three spatially dominant texture patches (i.e., the reference image is broken up into indexable pieces based on triads of contiguous texture defined regions). The similarity ratio is calculated for the entry (i.e., based on the texture patch centroids) and placed in the appropriate bins, along with its corresponding texture descriptor (3 patches x 16 multiscale oriented energies = 48 dimensional unit vector, with normalization performed across all dimensions), i.e.,

$$\hat{D}(p,\theta,\sigma) = \frac{D(p,\theta,\sigma)}{\|D(p,\theta,\sigma)\|} \tag{4}$$

where $\hat{D}$ and D represent the normalized and non-normalized feature vectors respectively, with

$$\|D(p,\theta,\sigma)\| = \sqrt{\sum_{p' \in \{patches\}} \sum_{\theta' \in \{orientations\}} \sum_{\sigma' \in \{scales\}} D(p',\theta',\sigma')} \tag{5}$$

where p is the patch, θ is the orientation and σ is the scale. At this stage it is expected that each bin will have more than one reference image database entry; however, texture is in the ready to support refined distinctions.

Given a video, texture patches are extracted in the same fashion as in building the database entries. Three contiguous texture patches suffice to define a probe to match against the database. The probe similarity ratio for the video texture patches restricts consideration to database entries in the bins that cover the probe value. Thus, coarsely quantized geometry has served as an initial filter on matching.

Typically, more than one database entry will populate each similarity ratio bin. The final match is established by selecting the most similar database entry in the indexed bin through consideration of appearance. Appearance similarity is quantified by calculating the inner product between the video and database texture descriptor vectors, (i.e., given that the vectors have been normalized, by calculating the cosine between their directions); values closer to unity indicate greater similarity.

2.5 Preliminary Experiment

In our preliminary experiment we generated a set of database entries and a set of probes to evaluate empirically the proposed approach. To generate a probe or a database entry we manually specified 3 homogeneous texture patches that are spatially nearby. Hence, we are simulating an automatic segmentation that creates a database or probe entry. Following selection, each database entry and probe has its respective geometry (similarity ratio) and its appearance (48 dimensional appearance vector) calculated through fully automated means.

The data set that we used derives from one real world orthoimage, Mazsea, captured from Wisconsin, see Figure 3. Significantly, Mazsea has an interesting variety of texture that is typical of aerial imagery, including forest, city and agricultural. The derived database consisted of 8 entries extracted from Mazsea, see Figure 4. Each database entry consists of 3 texture patches and their geometries. Based on the geometry each database entry was assigned to its respective bin. The probe set also was of cardinality 8. Here, we simulated flyovers of Mazsea by reselecting 3 texture patches along the linear spans used to generate database entries illustrated in Figure 4. All-way matches were performed between the probes and database entries under the proposed approach to combined appearance/geometry representation and matching. Results are reported as a confusion matrix, Table 1. It is seen that no confusions were made in this experiment. Moreover, correct matches are well separated from runner-ups. These results lend empirical support to the proposed approach.

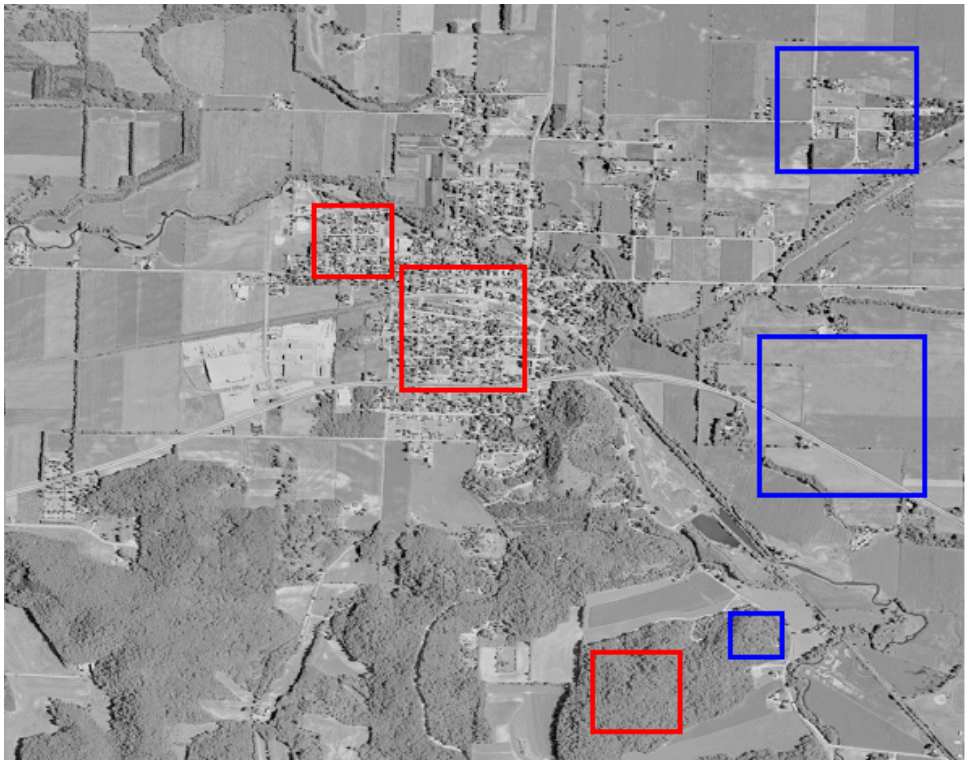

Figure 3: Mazsea Imagery. Our preliminary experiment made use of the depicted orthophoto to derive both database entries and probes. Highlighted regions illustrated typically selected texture patches.

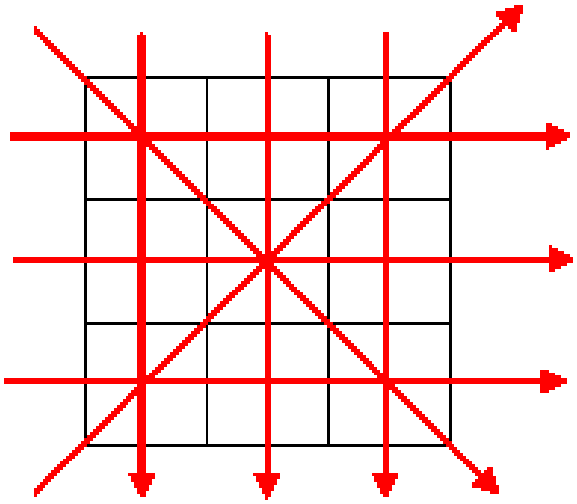

Figure 4: Schematic Representation of Strips Used to Generate Database Entries and Probes in Preliminary Experiment. Eight entries/probes were created along the lines indicated.

models \ probes							
	0.975106						
		0.971647				0.035413	0.674236
			0.938074		0.467941		0.484289
	0.557313			0.932360			
			0.408941		0.984367	0.076012	
			0.377967		0.073822	0.988639	
		0.063635				0.971892	0.230533
		0.649547				0.250322	0.980091

Table 1: Confusion Matrix Showing Results of Preliminary Experiment. Probes are shown along the horizontal; database models along the vertical. Perfect performance was achieved as every probe showed the best (highest) match when compared to the corresponding database model (i.e., along the diagonal). Numbers indicated the inner product for each comparison between a probe and database entry appearance vector. Blank entries arise for cases where the geometric prefilter pruned consideration of the database entry; so, appearance was not considered. Note that the matrix is not expected to be symmetric as texture patches were selected differently for probes and database entries.

2.6 Survival of the fittest

It is important to note that other approaches for representing appearance, geometry and their combination were investigated; however, during our empirical evaluation the approach described above was the clear winner. Alternative approaches considered are outlined in the appendices to this report.

3 Refinements

To make the proposed approach more generally applicable to the challenges outlined in the introduction it must be refined. Toward this end, this section of the report presents methods for making the appearance representation invariant to rotation and scale as well as for extending the geometric representation to encompass two-dimensional geometry.

3.1 Rotational invariance

One important aspect that we need to deal with is invariance to rotation. Oriented texture appears differently under in plane rotations. For us, it is important to be robust to rotation between the probes and the database entries as it would correspond to view changes in the sensor platforms.

The solution to this problem lies in the fact that rotation space is cyclic, e.g., rotating the image by 360 degrees makes no change. We can exploit this fact to make our representation robust to rotation, in other words we can align two representations in rotation space. Our oriented energy image representation coarsely samples the orientation space into 0, 45, 90, 135 degrees (although the underlying filters have a cosine tuning in bandwidth and therefore are fairly broadly tuned about their peak central frequencies). In order to make this representation invariant to rotation, we align the images so that their dominant orientations match: Assume we have a pair of 4 dimensional vectors, $v = (O_1, O_2, O_3, O_4)$ and $u = (O_1, O_2, O_3, O_4)$,O_i where O_i represents the mean energy at a particular orientation, and orientations are 0, 45, 90, 135 degrees. Define cyclic permutation $v(p)^p$ as:

$$v(p)^p = (O_2, O_3, O_4, O_1). \tag{6}$$

Clearly there are 4 different permutations of a 4 dimensional vector. The permutation that best aligns the two vectors, u and v, is taken as

$$\arg\max_k \{(v)^{p^{(k)}} \cdot u\} \tag{7}$$

where $\cdot$ is the inner product, k is the number of times the cyclic permutation is made so that $p^{(k)}$ denotes the k^{th} permutation. Thus, the aligning permutation is the one that maximizes the match. This procedure is performed at each scale separately. So, having S different scales

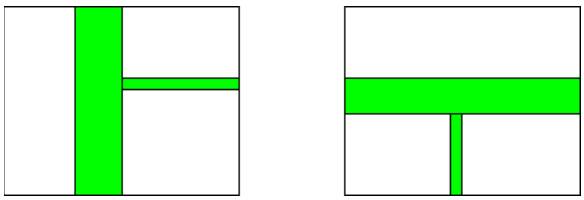

Figure 5: Rotational Invariance through Alignment of Orientation Distributions. The left and right panels schematize the same image rotated by 90 degrees. There are 4 possible permutations when aligning the right image to the left (given that we quantize to 4 orientations, horizontal, vertical and two diagonals). In general, one of these permutations will yield the highest match score as it correctly aligns the orientation distributions. In the example shown, the highest match score will be achieved when the right energy image representation is rotated counter clock wise by 90 degrees. Corresponding segments would be aligned and rotational invariance would be achieved.

we could get S distinct k values. We decide on a global permutation k for the whole patch descriptor based on the maximal number of votes for each particular permutation. In other words if at scale0, scale1, scale2, scale3 the best match permutations respectively are 2, 3, 1, 2 then the whole patch would be permuted twice at each scale because 2 had maximal support. We break ties by always choosing a smaller number of permutations arbitrarily. Following calculation of the permutation that best aligns two vectors of interest, subsequent match scores are calculated correspondingly. An example of this approach to alignment is given in Figure 5.

Performance of the above mentioned method under rotation is summarized in Figure 6. The experiment is the same as the preliminary experiment described in the previous section, except now the queries are selected from a rotated image. It is important to note that deterioration due to rotation is purely due to appearance, the geometric descriptor (similarity ratio) is completely invariant to rotational changes by construction. Without the proposed method for realizing rotational invariance, it is evident that performance decreases with an increase of the angle of rotation. As we rotate the queries by 90 degrees the performance of the algorithm drops to 25% correct. Once the proposed method for rotational invariance is included in the processing, performance returns to 100% accuracy across the range 0 to 90 degrees rotation.

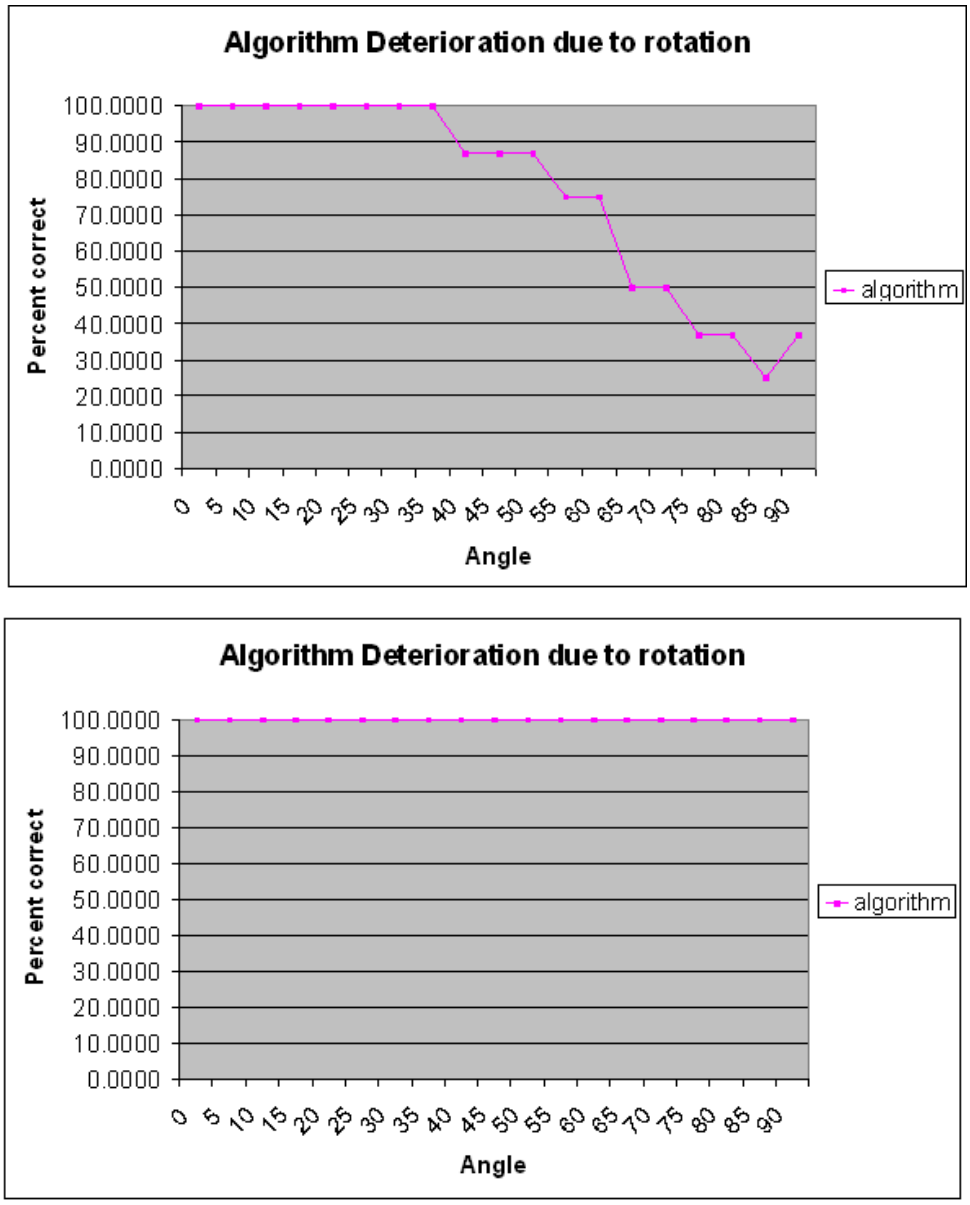

Figure 6: Match Accuracy Before and After Compensation for Rotation. The top panel shows percent correct as a function of rotation prior to compensating for rotation. The bottom panel shows percent correct with rotation compensation. The comparison data is as with Figure 3, except now individual texture patches are systematically rotated between the probes and reference database.

3.2 Scale invariance

A second important challenge that we need to face is scale invariance. Texture looks different at different scales; see, e.g., [20] for a good general discussion of scale, its importance and scale invariance. A city or a forest would look drastically different as seen from a satellite compared to the way it looks from a low flying aircraft. While, it is possible to compensate for scale based on a priori knowledge (e.g., regarding relative acquisition altitude and sensor resolution of video vs. reference), here we seek an image-based approach in keeping with the constraint that limited telemetry is available.

The proposed approach to scale invariance is very similar to the proposed approach to rotation invariance, except for the fact that scale is not cyclic, i.e., systematic scale change never returns to the starting scale. Still, an approach can be realized based on the notion of aligning energy across scale (rather than orientation). For this purpose, we collapse across orientation (i.e., sum) and get a set of integrated measures of energy as a function of scale: At each scale i collapsing across orientation is formulated as:

$$S_i = \sum_{j \in \{orientations\}} E_{\sigma_i \theta_j} \tag{8}$$

Where j is the number of different orientations S_i is the amount of energy at scale i and $E_{\sigma_i \theta_j}$ is the energy at scale σ_i and orientation θ_j. The cardinality of the derived set is equal to the number of scales in the underlying oriented energy image representation, 4 in the case of the present study (see Section 2.2).

Given two (collapsed across orientation) texture vectors to be matched, we shift their representations in scale so that the energy distributions align best under the inner product norm. Formally, assume we have a pair of 4 dimensional vectors: $v = \{S_0^1, S_1^1, S_2^1, S_3^1\}$ and $u = \{S_0^2, S_1^2, S_2^2, S_3^2\}$ where S_i^j is the energy at scale i of vector j, we take

$$\begin{aligned} \max\{&(S_0^1, S_1^1, S_2^1, S_3^1) \cdot (S_0^2, S_1^2, S_2^2, S_3^2), \\ &(S_1^1, S_2^1, S_3^1) \cdot (S_0^2, S_1^2, S_2^2), \\ &(S_0^1, S_1^1, S_2^1) \cdot (S_1^2, S_2^2, S_3^2)\} \end{aligned} \tag{9}$$

as indicating the best alignment of the representation across scale and calculate subsequent match scores correspondingly. For current purposes, scale shifts by a factor of 2 in both up- and down-scales are accommodated in the matching algorithm. This limitation was imposed because further shifts would yield too few common components for matching as extremal scales fall-outside the dynamic range of our scale space sampling. The essential ideas are elaborated in Figure 7. As illustrated in the figure, owing to the non-cyclic nature of scale-space, the shifting of one (finite) representation relative to another yields a subset of scales that have no corresponding component for matching in the other representation, i.e., extremal scales will shift outside the effective dynamic range of the representation.

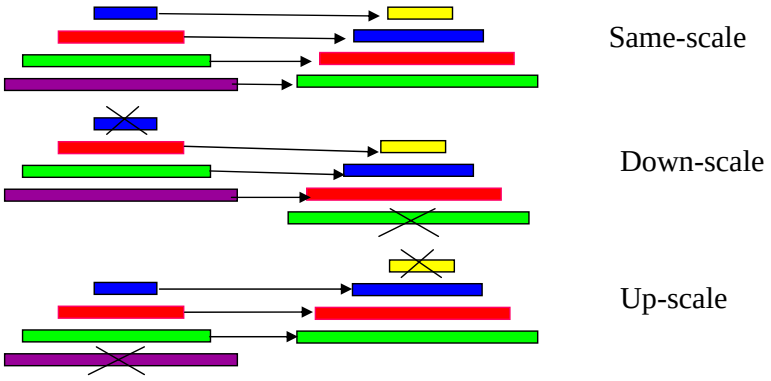

Figure 7: Scale Invariance through Alignment of Scales. The left and right columns schematize two identical patterns, except that they have been imaged at different scales. Each set of four rectangles in a pyramid-like arrangement corresponds to an energy image representation of the pattern (collapsed across orientation); smaller rectangles correspond to coarser scales. Between pyramids the colours are to be aligned as they correspond to the same physical attribute, albeit imaged at different scales. In the top row, due to differences in acquisition the corresponding scales appear at different levels in the pyramids. In the second row, a "down-scale" of the left representation relative to the right fails to align across scales. In the third row, an "up-scale" of the left representation relative to the right correctly aligns the scales. Due to finite resolution in the recovered representations, down- and up-scales necessarily yield pyramid levels that can no longer be compared between the two samples, as depicted with X's in the diagrams.

Performance of the proposed method for scale invariance is illustrated in Figure 8. The experiment is the same as that for rotational invariance (Section 3.1), except now the queries are selected from scaled images. Without compensation for scale, a severe degradation of performance is evident beyond 50% scale change. It is important to note that this deterioration due to scale arises purely from appearance; the geometric descriptor (similarity ratio) is completely invariant to scale changes by construction. In comparison, once the proposed method for adjusting for scale is included in the processing, the performance does not drop below 87% correct for the range of scales considered. It is seen that even under scale adaptation, errors do occur. The observed errors arise as the algorithm is capable of searching across a wider range of possibilities. In particular, problems arise as scales that contain the most discriminative information for a given pattern are shifted outside the dynamic range of the representation. For example, if fine scale (i.e., high spatial frequency) information serves to distinguish two patterns, but is dropped from consideration during the shift to align dominant scales, then the two patterns may be confused.

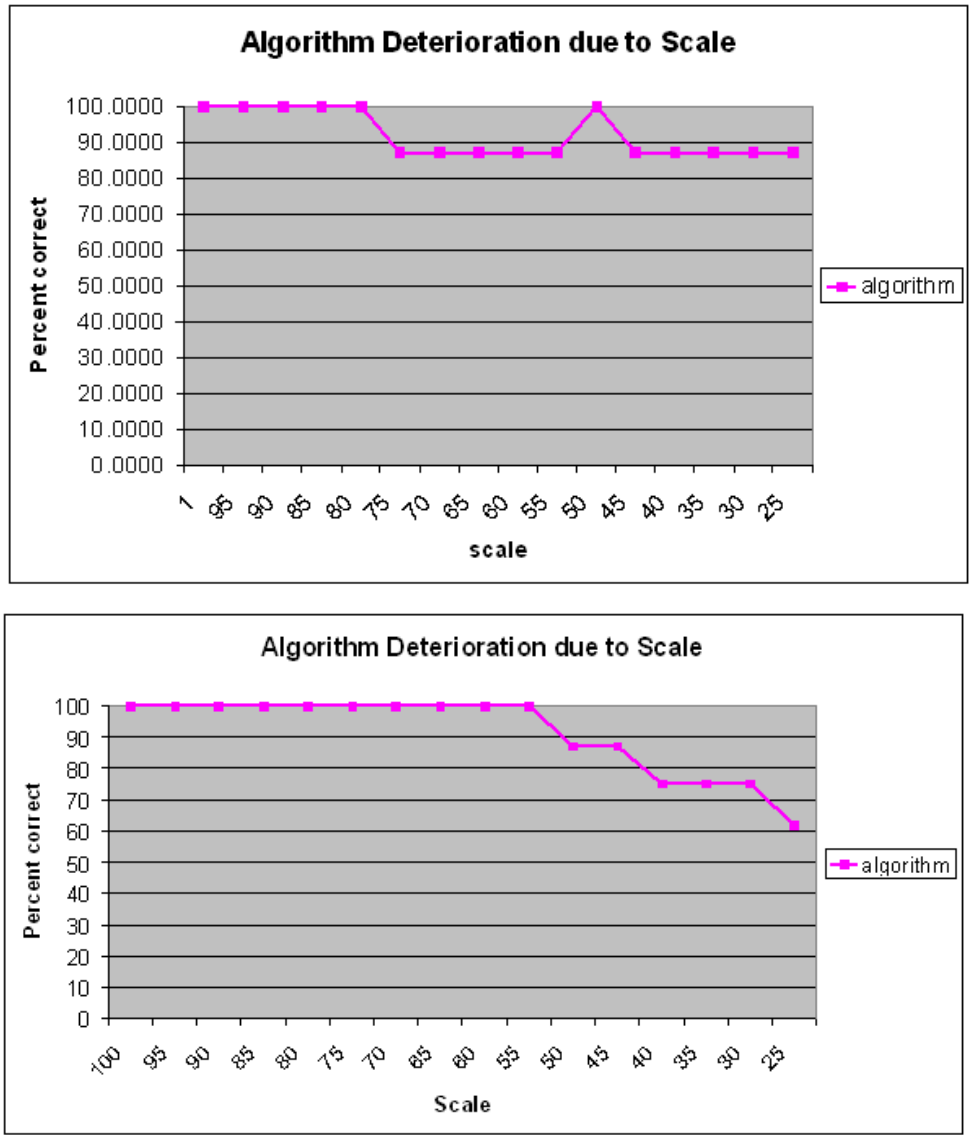

Figure 8: Match Accuracy Before and After Compensation for Scale. The top panel shows percent correct as a function of scale prior to compensating for scale. The bottom panel shows percent correct with scale compensation. The comparison data is as with Figure 3, except now individual texture patches are systematically scaled between the probes and reference database.

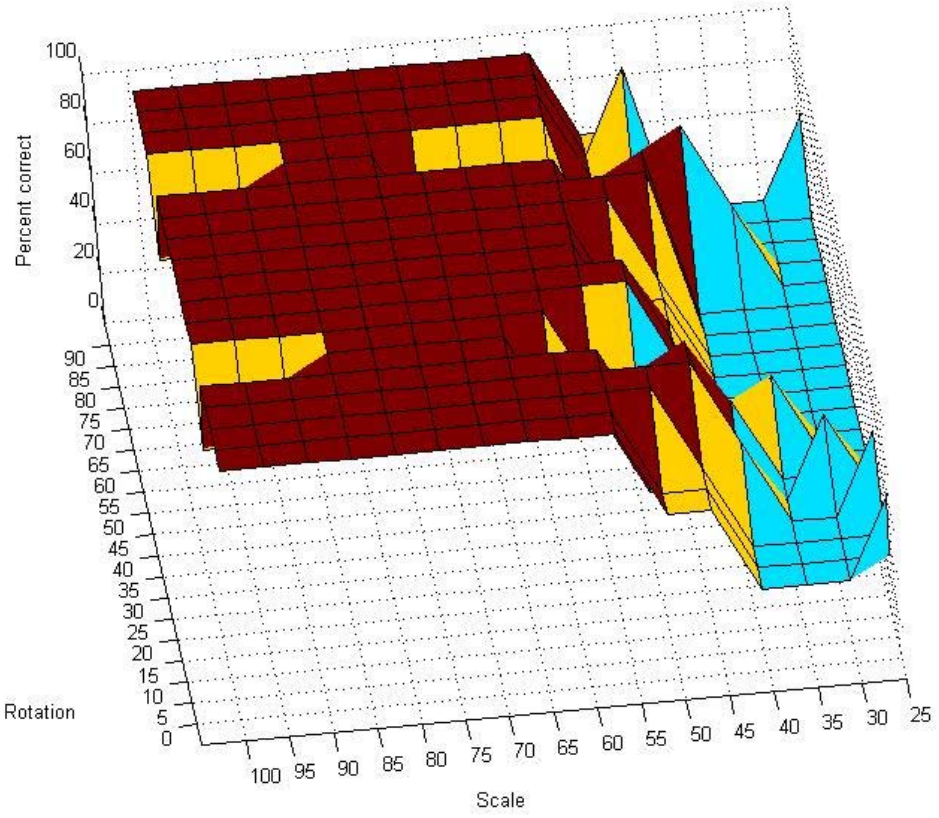

Figure 9: Match Accuracy as a Function of Simultaneous Rotation and Scale Changes.

3.3 Rotation and scale invariance combined

In practice, matching between probes and the reference database must be robust to changes in both rotation and scale when they occur simultaneously. To evaluate the effectiveness of the proposed refinements to simultaneous changes in scale and rotation, we conducted a combined experiment where patterns are first aligned for scale and then for orientation. It is appropriate to perform alignment in this order to make sure that only those scales mutually apparent in both of the image regions to be compared are brought into consideration for rotational analysis. Figure 9 shows the corresponding empirical results. In this case, the probe image data was manipulated to provide for variation in both scale and orientation. Strong performance is observed out to 50% scale changes, even as rotation varies across 90 degrees. Further scale changes remove too much discriminatory information and performance degrades.

3.4 2D geometry

In our preliminary approach, we exploited only one-dimensional geometric relationships between texture-defined image regions. In general, our image data is observed in the two-dimensional image plane. Correspondingly we now generalize our approach to encompass two-dimensional geometric relationships between image regions.

While it would be possible to directly generalize our one-dimensional similarity ratio to two-dimensions (and thus maintain invariance to the similarity group of transformations), we instead take this opportunity to extend the representational power to encompass affine transformations. Analytically, affine transformations capture the motion of a plane under orthography [17]. In practice, affine transformations of an image have been widely used in computer vision to compensate for viewpoint changes in cases where the relief of an object is small compared to the viewing distance [27], such a model is especially popular in the analysis of aerial imagery [18].

For current concerns, matters can be formalized with reference to a set of four image regions and their centroids. Three of the region centroids are used to establish a two-dimensional basis; the coordinates of the fourth centroid are projected on the basis to achieve affine invariance, see Figure 10 (c.f., [19]). More formally, let the image coordinates of the centroids of the four regions be given by P_0, P_1, P_2, P_3 then arbitrarily selecting the first three points as the basis set, the affine invariant coordinates of the fourth point, P_3^{aff}, are specified as $P_3^{aff} = P_0 + a(P_1 - P_0) + b(P_2 - P_0) = P_0 + aX + bY$, with a and b the orthogonal projections of P_3 on $X = (P_1 - P_0)$, $Y = (P_2 - P_0)$. Notice that given a basis set defined in terms of 3 regions, it is possible to represent an arbitrary number of additional regions in terms of the established basis set. In our current work we simply select a single additional region to investigate the discriminatory power of a small set of regions; thus, the relative geometry of the four patches is captured by the pair of numbers, a and b.

Working with a set of four image regions leads to twelve different ways to specify a basis set (i.e., 12 ways to choose 3 distinct items from a set of 4) for specifying the affine invariant coordinates of the remaining region. Following previous work [19], we build our reference image database with redundancy, so that for each model (i.e., reference image represented by 4 texture patches) we create 12 database entries, one for each choice of basis points. Subsequently, given a video probe characterized by four texture patches, any three patches can be selected to define the basis for the pair, (a, b), that is to be compared to the database: All possible choices are represented in the database. In practice, we select the set of three centroid coordinates that most nearly yields an orthogonal frame as it will be relatively resilient to centroid localization errors [13].

To combine geometry with appearance, we follow a straightforward generalization of the approach used in our preliminary approach: We use geometry as a prefilter for appearance-based matching. In particular, coarse quantization of the geometric parameters, (a, b), are used to define a two-dimensional look-up table. Each cell in the table is filled with the appearance

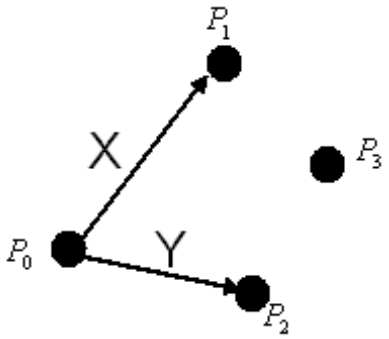

Figure 10: Affine Invariant Representation of Two-Dimensional Geometry. Given the centroids of 4 texture defined image regions, we select 3 to form a basis set of two vectors (X and Y). The centroid of the fourth region is projected on the basis vectors to yield a coordinate pair that is invariant to affine transformations of the plane.

vectors (i.e., oriented energy vectors) for reference images whose derived geometric parameters are covered by the cells range. During matching, a probe indexes a cell based on its derived geometric parameters; subsequently, the best match is taken as that which yields the smallest inner product between the appearance vector of the probe and all appearance vectors contained in the indexed cell. In accordance with the refinements for appearance rotational and scale invariance that were introduced earlier in this section, matching between candidate appearance vectors takes care to align energy distributions in scale and rotation.

4 Final experiment

In this section we describe an experiment that evaluates the effectiveness of the refined methods for matching video probes to a reference image database. Individual database entries and probes are characterized in terms of quadruples of manually selected, texture-defined image regions. The appearance of individual patches is captured via their 16 dimensional oriented energy vectors, including considerations for rotational and scale invariance in matching. The relative geometry of patches is captured via their 2 dimensional affine invariants.

4.1 Database

Our database had 7 entries corresponding to distinct geographical locations. Each entry was derived from a visible, 8-bit greylevel orthoimage at 1 ground meter/pixel spatial resolution. Four entries were from a region around Camp Lejeune, North Carolina, one entry was from Mazsea, Wisconsin (same region as in our preliminary experiment), one entry was from Jack-

sonville, Florida and one entry was from Sudbury, Ontario. We uniformly quantized the affine parameters into bins of size 0.5, with 0.2 overlap between adjacent bins. These units were chosen based on empirical inspection to ensure that geometric-based match pruning would run the gambit from providing complete disambiguation of match (appearance need not even be considered) to no disambiguation of match (appearance does all the work). Figure 11 provides a visualization of the resulting space. Note that many bins lie empty, which suggests that an adaptive binning strategy should be considered in future work.

4.2 Probes

Our probes derive from 5 distinct geographical locations that corresponded to 5 of the database entries. Probes 1 and 2 were constructed from real flyover videos of Camp LeJeune that share 50% overlap. The image resolution in these flyovers is approximately 0.7 ground meters/pixel; obliquity is approximately 45 degrees off the horizon. Probe 3 was constructed from the Mazsea orthoimage, with image warping to yield 50% scale change, 40 degrees rotation and shear relative to the reference image. Probes 4 and 5 were derived from the Jacksonville and Sudbury orthoimages, in the same manner.

4.3 Examples

The first 2 database entries are illustrated in Figure 12; their corresponding probes are shown in Figure 13. Both queries and probes have 50% overlap. Two additional database entries were derived from surrounding areas of Lejeune (not shown) to test further the ability of the proposed method to distinguish between nearby areas; however, no real flyovers were available for these regions. Database entries and probes 3, 4, 5 are shown in Figures 14, 15, 16, 17, 18, 19, respectively.

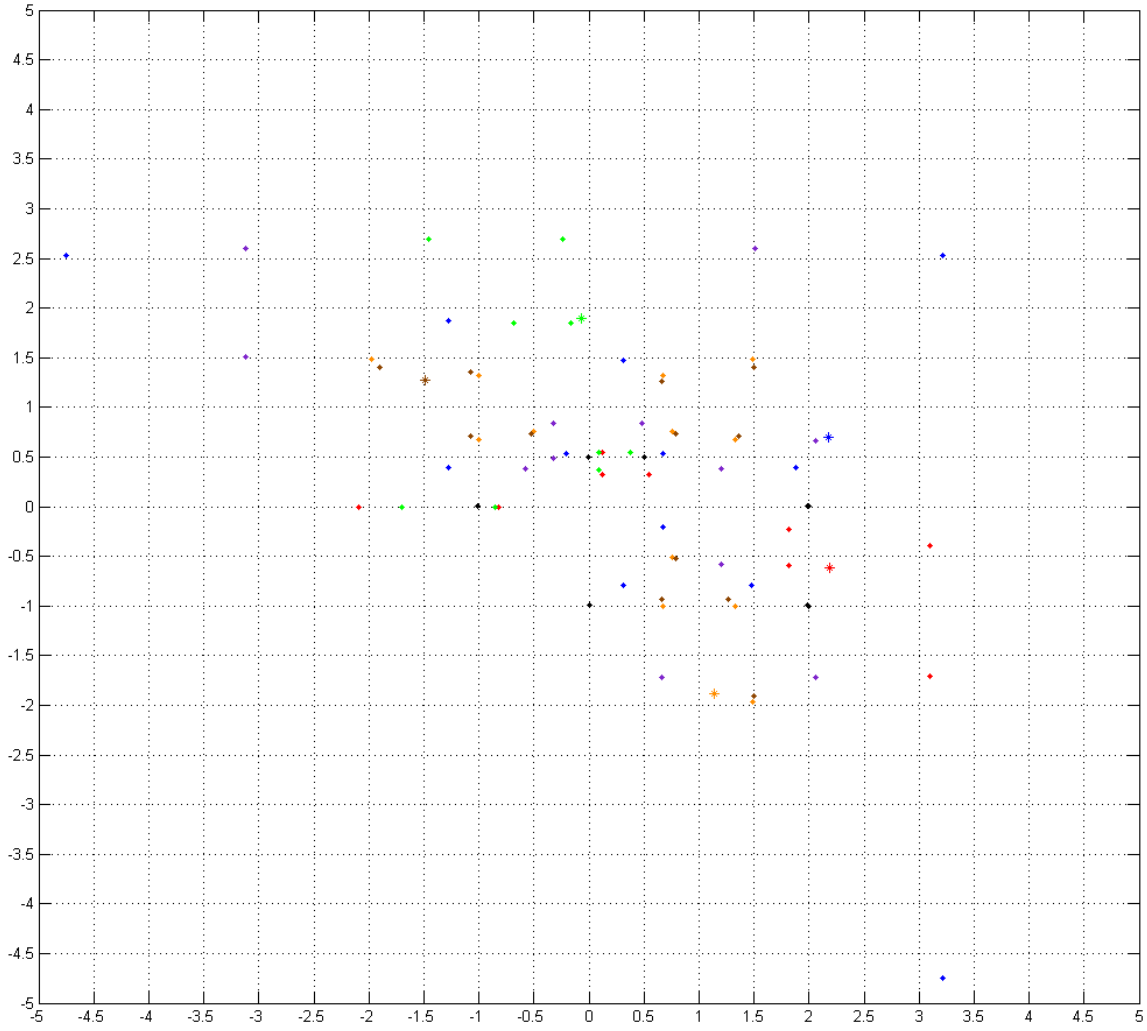

Figure 11: Depiction of Final Experiment Database Geometry. The axes show quantized affine geometry parameters. Dots correspond to populated cells. Each distinct model (i.e., geographic location) is symbolized by its own colour. Each model has multiple entries arising from our redundant approach to representing geometry, see Section 3.4

Figure 12: Database Entries 1 and 2. The first two database entries in our experiment were derived from the depicted orthophoto of Camp LeJeune, NC. Image regions defining entry 1 are highlighted in red. Image regions defining entry 2 are highlighted in green.

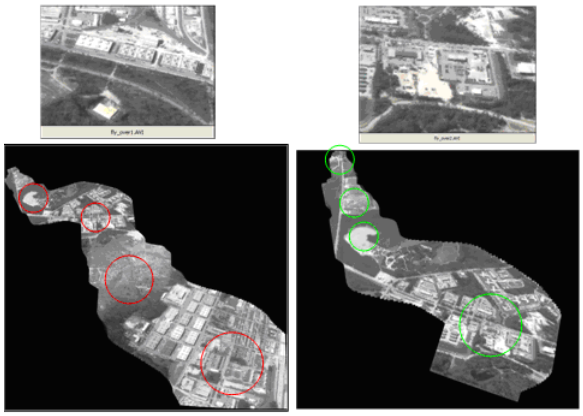

Figure 13: Aerial Video Used to Construct Probes 1 and 2. Single frames from two real flyovers are shown in the top row. Corresponding image mosaics encompassing the entire videos are shown below. Red and green highlight probe defining texture patches, analogous to Figure 12.

Figure 14: Database Entry 3 from Mazsea. Red highlights the defining texture patches.

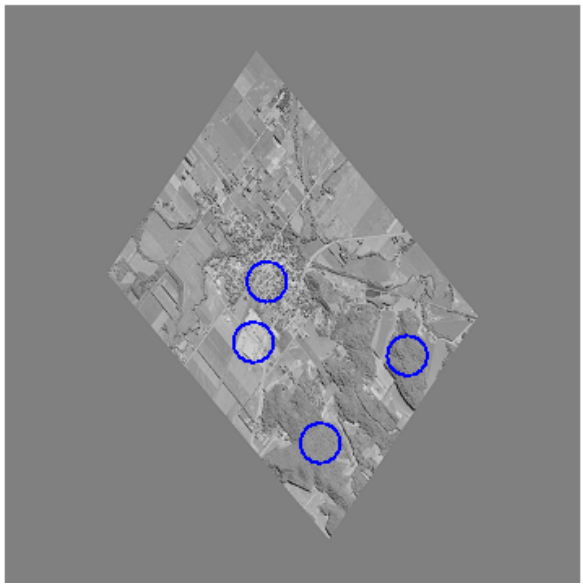

Figure 15: Probe 3 from Mazsea. Blue highlights the defining texture patches.

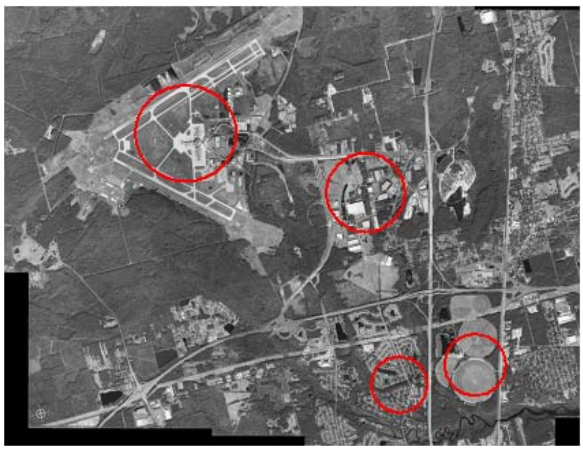

Figure 16: Database Entry 4 from Jacksonville. Red highlights the defining texture patches.

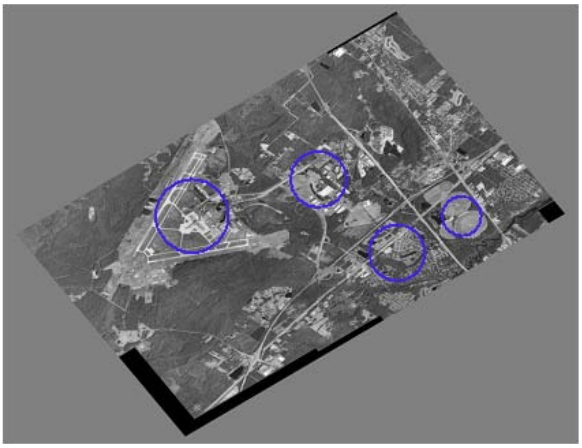

Figure 17: Probe 4 from Jacksonville. Blue highlights the defining texture patches.

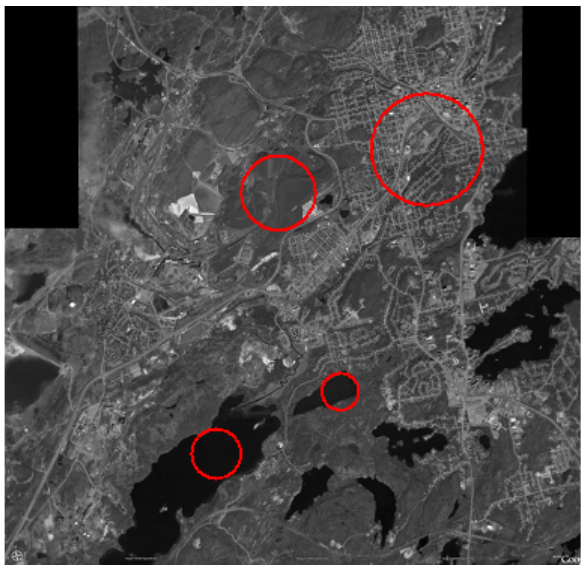

Figure 18: Database Entry 5 from Sudbury. Red highlights the defining texture patches.

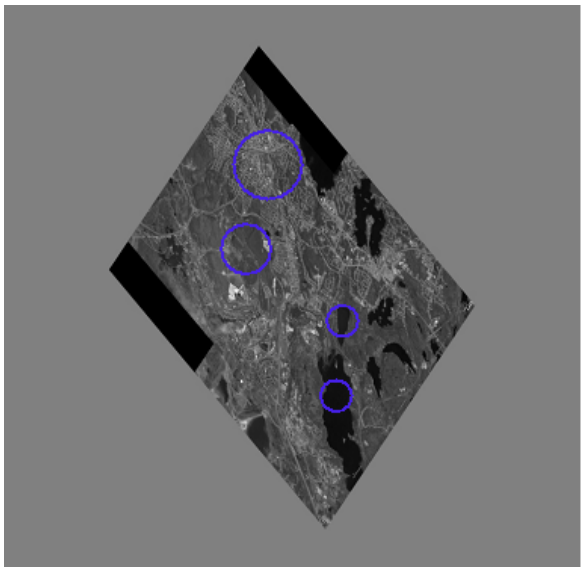

Figure 19: Probe 5 from Sudbury. Blue highlights the defining texture patches.

4.4 Matching methodology

Once populated, our database consisted of 7 x 12 entries (i.e., each of the 7 geographic regions contributed 12 redundant entries, see Section 3.4). For each of the 5 probes, we indexed into the geometry defined look-up table based on recovered affine parameters. For each entry in the indexed cell, we matched against all appearance vectors that were present using the inner product norm. Each appearance vector is of dimension 16 x 4 (16 oriented energies, 4 texture regions for each entry). The geographic location corresponding to the smallest computed inner product was declared as the final match.

4.5 Results

Results are summarized in Figure 20. In all cases, the probe is matched to the correct database entry. A subtlety occurs with respect to probe 1. In this case, the top match was for the correct database entry, but as captured by a different permutation of texture patches in construction of the affine basis. In all probes the number of database comparisons was greatly reduced, leading to a significant decrease in the number of appearance comparisons. Significantly, we also looked at the results of matching purely on the basis of appearance (i.e., ignoring the geometric prefilter) and found that mismatches occurred in this case. Overall, strong results are had only when appearance and geometry are combined.

5 Discussion

In this report we have presented a method for matching between aerial video and corresponding reference orthoimagery, as typical of geospatial databases. The method combines image appearance, characterized in terms of texture defined regions, and image geometry, characterized in terms of relationships between textured regions. By construction, the matching methods are robust to a range of photometric and geometric distortions between image sources, including changes in greylevel contrast and affine geometric transformations. Empirical investigations suggest the promise of the approach.

A variety of future research directions should be pursued. First, it is desirable to automate further all aspects of processing. In this regard, it is particularly desirable to remove the need for manual selection of image regions of interest in both database construction and probe specification. Second, it would be interesting to investigate a more integrated way of combining appearance and geometry (i.e., in contrast to the current method, which operates sequentially on geometry and appearance). Third, it is important to refine further the overall approach to make it more directly applicable to indexing large image databases. Fourth, it is important to subject the method and all further work along these lines to additional empirical evaluation, especially further evaluation that employs real operational data.

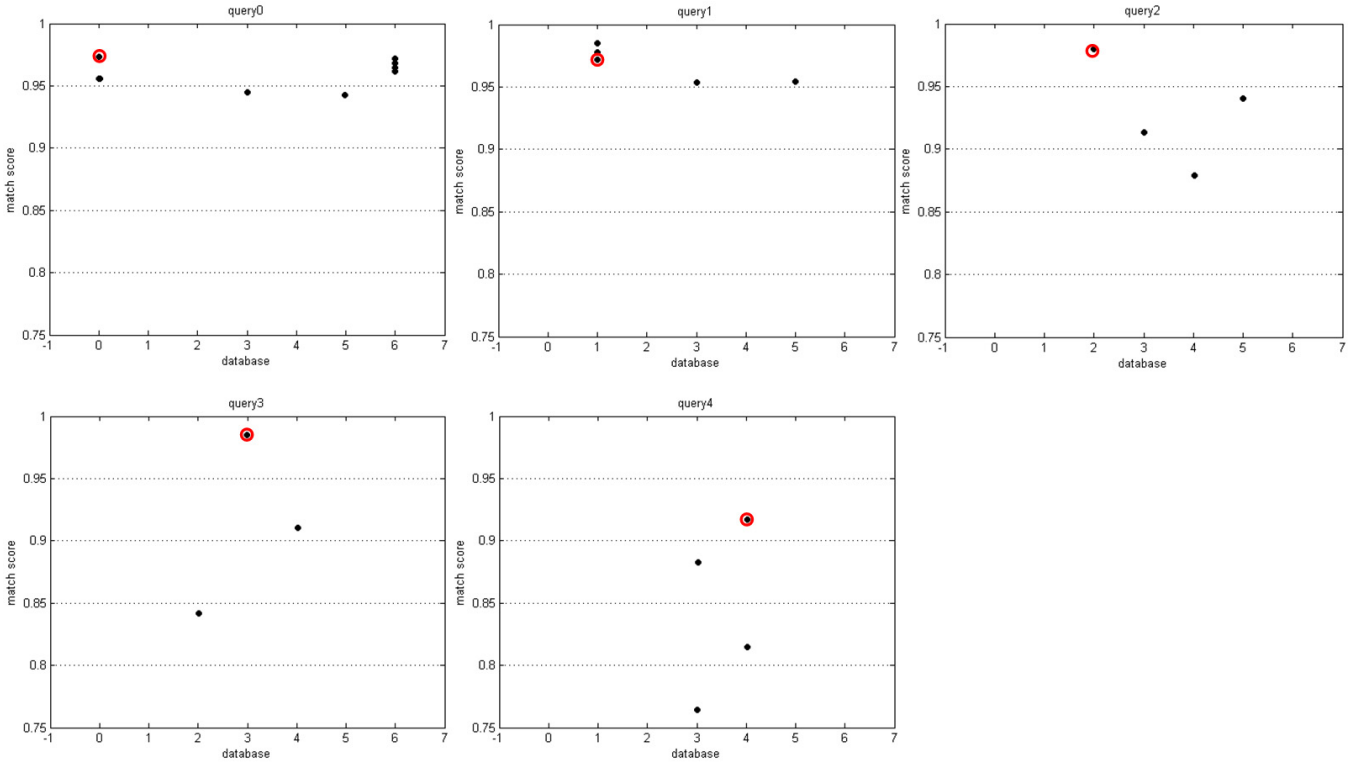

Figure 20: Results of Final Experiment. Each plot corresponds to the match results for a given probe. Within a plot, the abscissa indexes database entries; the ordinate shows final match scores. Due to redundancy in database specification, there can be more that one match/column. Red circles indicate matches between corresponding probes and database entries. See text for discussion.

A Alternative appearance and geometry representations

In addition to the methods illustrated in the main body of this paper we have considered various alternative representations for appearance, geometry and their combination. Below, we illustrate these approaches and the reasons for their failures.

A.1 Variance representation of appearance

One logical extension of using the mean to encapsulate the information available in each band of our multiscale, oriented energy representation of image appearance is to use higher moments of the distributions. To explore this possibility, we experimented with the use of variance, i.e., we calculated multiscale, oriented energy in exactly the same manner as described earlier in this report; however, to construct the final descriptor vectors we calculated the energy variance rather than the mean. Such an approach has potential to maintain additional information regarding how energy is distributed across a region of analysis in comparison to the mean statistic. Preliminary empirical investigations with this method showed that it had reduced discriminatory power in comparison to the means statistic. In particular, it seemed to differ little between different image textures at lower spatial frequencies. Further, attempts to align energy distribution, to achieve rotational and scale invariance proved to be less stable when using variances in comparison to means. These shortcomings arise due to decreased numerical stability in the calculation of the higher order moment (variance vs. mean). For these reasons, the variance representation was dropped from consideration.

A.2 Linear ordering representation of geometry

To explore the power of purely ordinal representations of geometry, we considered linear ordering of texture defined regions. As with the similarity ratio (Section 2.3), linear ordering was calculated by projecting three consecutive patches on the best least squares fit to a line. In this case, however, rather than using the projections to calculate a numerical quantity, the geometry was simply used to order the texture vectors as they were encountered. During empirical evaluation, it was found that this approach decreased the discriminatory power of matching relative to the use of the similarity ratio; therefore, it was dropped from further consideration. Apparently, this relative performance was due to the fact that for the range of geometric changes that were studied between database and probe entries, the similarity ratio (and subsequently the affine parameters) provided an adequate model. If in future research it is found that the similarity and affine invariants do not capture the observed range of geometric variation (e.g., due to consideration of a wider range of viewpoints), then it will be appropriate to reconsider alternative geometric characterizations. For example, projective invariants [1], various qualitative measures (e.g., ordinal relationships) and interval analysis [24] methods could be considered.

A.3 Alternative appearance/geometry combinations: Appearance as prefilter for geometry

In addition to using geometry as a prefilter for appearance-based matching, we also have considered use of appearance as a prefilter for geometric-based matching. To realize such

an approach, we employed a coarsely quantized appearance space based on clustering in the space of appearance vectors. Clustering was performed via k-means analysis [7]. Here, initial matching was performed in the (clustered) appearance space, with subsequent geometric-based matching used to distinguish between entries that mapped to the same appearance defined cluster. In particular, given a triad of image regions with associated appearance vectors and similarity ratio, appearance-based prefiltering was performed by selecting the nearest cluster. Within a cluster, the final match was taken as that which minimized the absolute difference between the similarity ratio of the probe and all initially indexed database entries.

Use of appearance as a prefilter for geometry-based matching was abandoned because of the implications of making our approach invariant to scale had for clustering. In particular, when adapting appearance vectors for scale, their dimensionality can be reduced on the fly, which makes it questionable to make use of clustering that has been predefined based on the range of scales present in reference image database entries only. A potential way to avoid such problems is to build the database appearance-based clusters redundantly with respect to available scales, e.g., in analogy with geometric redundancy used to construct database when geometry is used as a prefilter.

An alternative approach to combining appearance and geometry that we did not consider in the present study was to strive for a more integrated measure, i.e., a measure that did not have one source of information take precedence over the other. Such an investigation should be considered in a subsequent study.

B Alternative comparison metrics

In addition to the method illustrated in the main body of this paper for comparing two appearance vectors, alternatives were considered. Below we illustrate these approaches and explain why we ultimately made use of the inner product metric.

B.1 Inner product

For the sake of completeness, we begin by giving the definition of the inner product that we made use of in the main body of this paper. Given two n-dimensional vectors, X and Y, the inner product is defined as

$$X \cdot Y = \sum_{i=1}^{n} x_i y_i \tag{10}$$

where x_i and y_i are the individual vector components. We considered the inner product because it quantifies the degree of angular separation between two unit vectors and our texture descriptors are normalized by definition.

B.2 Euclidean distance

The Euclidean distance for two n-dimensional vectors, X and Y, is defined as

$$Euclidean(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{11}$$

where x_i and y_i are individual components. We consider Euclidean distance as it is arguably the most often employed distance metric.

B.3 Earth movers distance

The Earth Movers Distance (EMD) is a distance between 2 distributions defined as the minimal amount of work necessary to transform one distribution to the other [25]. Imagine one distribution as earth situated on a grid, now imagine a second grid where there are holes corresponding to the second distribution. The EMD is then the least amount of work needed to transfer the earth to the holes. This metric is useful because it is possible to define a custom ground distance matrix that captures the difficulty of moving earth from bin A in first distribution to bin B in the second distribution. Depending on your knowledge of the data it is possible to define a distance matrix that accentuates the desired similarities between the distributions. Given two n-dimensional vectors, X and Y, the EMD is defined as

$$EMD(X,Y) = \frac{Work(X^t, Y^t, F)}{\sum F} \tag{12}$$

where X^t, Y^t are the remappings of the original 1D feature vectors to 2D distributions parameterized in terms of scale and orientation and F is the difficulty of moving (flow) between any pair of elements in the scale $\times$ orientation matrix.

In the absence of any other knowledge, we took flow as proportional to absolute distance in scale and cyclic (absolute) distance in orientation. In particular, let scale, σ, and orientation, θ, range over four values, that we specify as 0, 1, 2, 3. Then the distance matrix for 2 distinct locations in the distribution is defined as follows.

$$\begin{aligned} dist\{(\sigma_1,\theta_1),(\sigma_2,\theta_2)\} &= abs(\sigma_1 - \sigma_2) + abs(\theta_1 - \theta_2), && \text{if abs}(\theta_1 - \theta_2) < 3 \\ dist\{(\sigma_1,\theta_1),(\sigma_2,\theta_2)\} &= abs(\sigma_1 - \sigma_2) + 1, && \text{if abs}(\theta_1 - \theta_2) = 3 \end{aligned}$$

We considered the EMD due to recent success with using this metric in a texture-based image indexing application [25].

B.4 Robust metric

The robust metric makes use of the inner product, but instead of considering one feature vector it considers the best inner product of all possible subsets of the feature vectors of size n-k,

with k the number of features we ignore during subset construction, similar to the RANSAC matching scheme [8]. In particular, given two n-dimensional vectors, X and Y, our robust metric is defined as

$$Robust(X,Y) = \max_{\forall(x,y) subset\_of\_size\_n-k(X,Y)} (x \cdot y) \tag{13}$$

Note that we have n choose $n-k$ different subsets. We consider a robust match metric as it allows a degree of tolerance to outlier contaminated feature vectors.

B.5 Relative performance of comparison metrics

Prior to empirical comparison of the various metrics it of interest to consider their relative dynamic ranges. Assuming that X and Y are normalized positive vectors ($X \cdot X = Y \cdot Y = 1$), the dynamic range of the inner product is $[0,1]$. The dynamic range of the Euclidean distance is $[0,\sqrt{2}]$ ($\sqrt{2}$ is the maximal distance between 2 positive normalized vectors) and it is inversely related to the inner product (if the inner product is 1, then Euclidean distance is 0). To map Euclidian distance to the same range as the inner product, we instead consider the quantity $1 - \frac{Euclidean(X,Y)}{\sqrt{2}}$. The robust metric has the same dynamic range as the inner product, $[0,1]$, by construction. Finally, the dynamic range of EMD is $[0,\infty]$, with an inverse relationship relative to the inner product, i.e., for the EMD the closer the distance is to zero the closer the items being compared. The unbounded nature of the EMD dynamic range makes it difficult to map to the unit interval, as we have for the other metrics under consideration.

For empirical comparison, texture vectors were extracted from the Mazsea image, Figure 3. For the database, 3 texture patches were selected corresponding to each of three types of ground cover, forest, urban and agriculture. For the queries, the same patches were selected, but the original image was distorted using a range of similarity transformations. The resulting percentages of correct matches, as functions of changes in scale and orientation, are shown in Figure 21 for each of the comparison metrics.

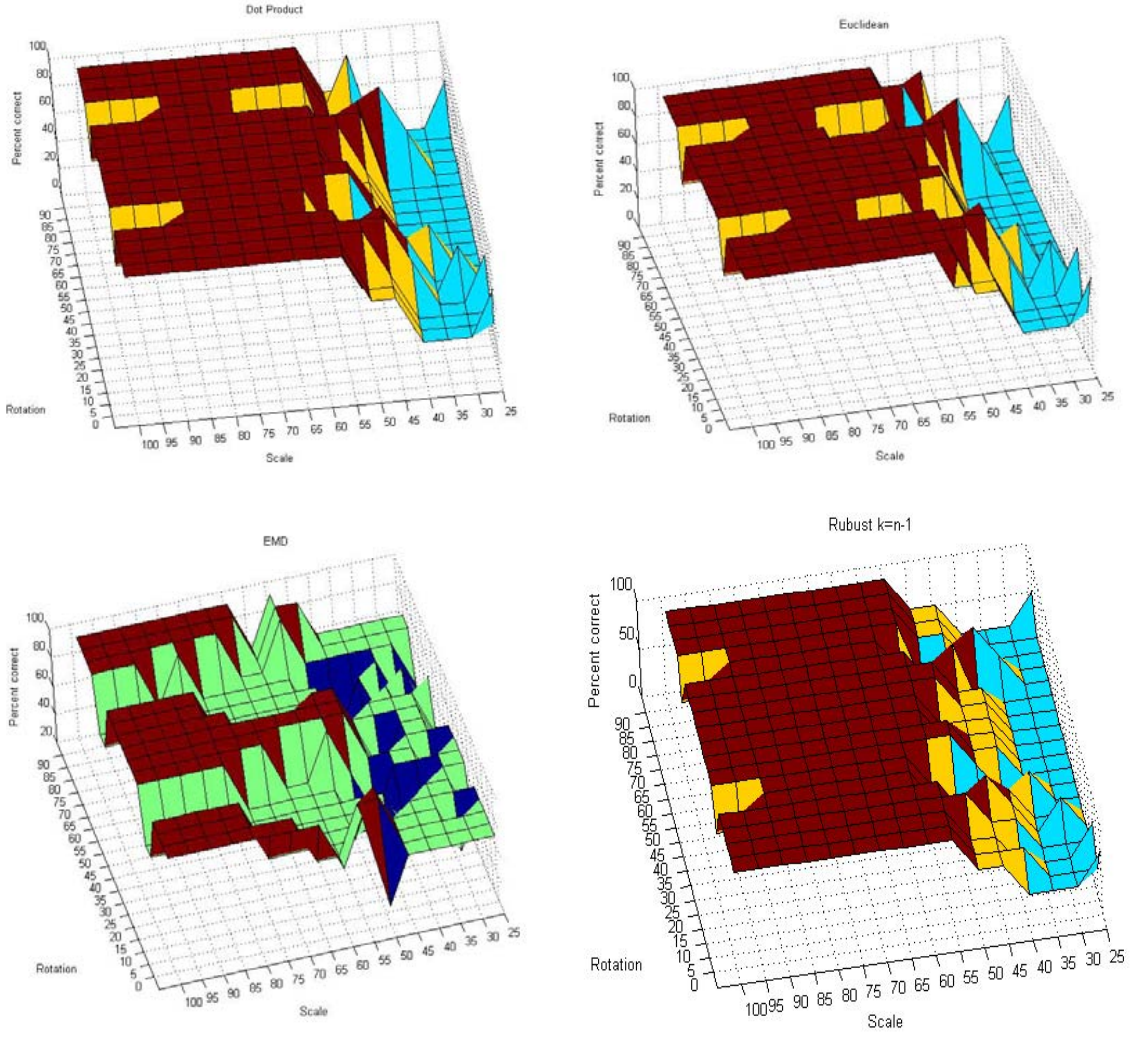

Figure 21: Results of Final Experiment. Each plot corresponds to the match results for a given probe. Within a plot, the abscissa indexes database entries; the ordinate shows final match scores. Due to redundancy in database specification, there can be more that one match/column. Red circles indicate matches between corresponding probes and database entries. See text for discussion.

Various observations are of significance. First, the results strongly indicate that the inner product and Euclidean distance are very similar in performance under the entire range of experimental settings. This was to be expected as the inner product is defined in terms of the cross-terms of the Euclidean distance and normalization of the texture vectors discounts any differences arising from the squares of the individual vector components. Given these observations, the inner product is preferable on the basis of simplicity. Second, even though the EMD performance never falls to 0 percent correct, it is inferior to the other 3 metrics, e.g., in having the same kinds of errors as the inner product, albeit at a wider range of scales and orientations. Apparently, the increased ability to match that is afforded by the EMD relative to the other metrics allows it to find low cost matches even when not appropriate. Third, the robust metric performs well. The depicted set of results arises from allowing the robust metric to drop at most one appearance component, as greater exclusion led to matching on too little data to support the desired distinctions. However, this metric turns out to yield a compressed range of match scores as poor match components are dropped in the search for the best match across all subsets of vector components. The compressed range may lead to problems in making distinctions in large databases. Still, in the future it might be appropriate to consider a robust metric, if outliers becomes of particular importance. Overall, these results led us to use the inner product metric based on simplicity and effectiveness.

References

[1] E.B. Barrett, P.M. Payton, N.N. Haag, and M.H. Brill. General methods for determining projective invariants in imagery. *Computer Vision Graphics and Image Processing*, 53(1):46–65, January 1991.

[2] R. Basri and D.W. Jacobs. Recognition using region correspondences. In *IEEE International Conference on Computer Vision*, pages 8–15, 1995.

[3] T.M. Breuel. Adaptive model base indexing. In *Proceedings of the Defence Advanced Research Projects Agency*, pages 805–814, 1989.

[4] L.G. Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376, December 1992.

[5] P.J. Burt. Smart sensing with a pyramid vision machine. *Proceedings for the IEEE*, 76(8):1006–1015, August 1988.

[6] S. Dickinson. Object representation and recognition. In E. Lepore and Z. Pylyshyn, editors, *What is cognitive science?*, chapter 6. Basil Blackwell Publishers, Malden MA, 1999.

[7] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, New York, 2001.

[8] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography.

[9] P.J. Flynn and A.K. Jain. 3d object recognition using invariant feature indexing of interpretation tables. *Computer Vision Graphics and Image Processing*, 55(2):119–129, March 1992.

[10] D.A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ, 2003.

[11] W.T. Freeman and E.H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, September 1991.

[12] V. Gaede and O. Gunther. Multidimensional access methods. *ACM Computing Survays*, 30(2):170–231, June 1998.

[13] W.E.L. Grimson, T. Lozano-Perez, and D.P. Huttenlocher. *Object Recognition by Computer: The Role of Geometric Constraints*. MIT Press, Cambridge, MA, 1990.

[14] R.M. Haralick. Statistical and structural approaches to texture. *Proceedings for the IEEE*, 67(5):786–804, May 1979.

[15] J. Illingworth and J.V. Kittler. A survey of the hough transform. *Computer Vision Graphics and Image Processing*, 44(1):87–116, October 1988.

[16] M. Irani and P. Anandan. Video indexing based on mosaic representations. *Proceedings for the IEEE*, 86(5):905–921, May 1998.

[17] F. Klein. *Elementary Mathematics from an Advanced Standpoint: Geometry*. Macmillan, NY, NY, 1939.

[18] R. Kumar, H. Sawhney, S. Samarasekera, S. Hsu, Tao H., Y. Guo, K. Hanna, A. Pope, R. Wildes, D. Hirvonen, M. Hansen, and P. Burt. Aerial video surveillance and exploitation. *Proceedings for the IEEE*, 89(10):1518–1539, October 2001.

[19] Y. Landman, J. Shwartz, and Wolfson H. Affine invariant model-based object recognition. *IEEE Transactions on Robotics and Automation*, 6(5):578–589, October 1990.

[20] T. Lindeberg. Scale-space: A framework for handling image structures at multiple scales. In *CERN School of Computing*, pages 8–21, 1996.

[21] P. Lipson, W.E.L. Grimson, and P. Sinha. Configuration based scene classification and image indexing. pages 1007–1013, 1997.

[22] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[23] W.J. MacLean and J.K. Tsotsos. Fast pattern recognition using gradient-descent search in an image pyramid. In *International Conference on Pattern Recognition*, pages 2873–2877, 2000.

[24] R. Moore. *Interval Analysis*. Prentice-Hall, Upper Saddle River, NJ, 1966.

[25] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *IEEE International Conference on Computer Vision*, pages 59–66, 1998.

[26] M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, November 1991.

[27] E. Trucco and A. Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, Upper Saddle River, NJ, 1998.

[28] R.P. Wildes, D.J. Hirvonen, S.C. Hsu, R. Kumar, W.B. Lehman, B. Matei, and W.Y. Zhao. Video georegistration: Algorithm and quantitative evaluation. In *IEEE International Conference on Computer Vision*, pages II: 343–350, 2001.

[29] K. Wong, E. Petrakis, and M. Spetsakis. Video georegistration: Algorithm and quantitative evaluation. In *Vision Interface*, pages 482–489, 1999.

[30] J. Zhang and T.N. Tan. Brief review of invariant texture analysis methods. *Pattern Recognition*, 35(3):735–747, March 2002.