A Method for Automatic Recognition of Small Mobile Targets in Aerial Images

Stuart Heinrich, Karthik Krish, Wesley Snyder, Siamak Khorram

Abstract—Vision based object recognition is usually done by matching morphological features, which are typically more invariant to lighting changes and more descriptive than color features. However, for small and/or distant objects extracting reliable shape information can be difficult or impossible, as sampling effects contribute more to image formation than the shapes of the objects being imaged, as the image gets smaller. For a given sensor, this limits the effective range that object recognition based on shape features is possible. Objects can also be recognized based on their color information, and this is more invariant to under sampling. In this paper, we are specifically interested in the problem of mobile recognition for a vision based guidance system. We show how a very small object previously identified in a survey image can be recognized by a vision based guidance system based on color information only, thereby allowing recognition to occur at much greater ranges than would be possible using a detector that used spatial queues.

Index Terms—Automatic target recognition, object recognition, registration, aerial, homography, projectivity, histogram shift.

1 Introduction

Object recognition under general conditions remains a very difficult task in computer vision, and the majority of progress in this area is still focused on the case where the objects to be recognized are large enough in the image for shape features to be utilized in the recognition process. In most cases, shape-based features such as gradients, curvature, corners, textures, etc are more reliable than color information which is highly dependent on view and lighting conditions. However, as sampling is reduced these morphological features become increasingly unreliable, making them unusable for recognition of objects that project to a very small area in the image.

Recognition is still possible at such small scales, but color features (which are more robust to sampling) must be used instead of morphological ones.

The local histogram is a good compact representation of the spatially independent color information in an image.

In most automatic target recognition tasks, targets are assumed to be close enough to the imaging device that they occupy a sufficient portion of image space in pixels for salient features of the target to be identified [Pop94] [Low04] [Low99] [BMP02], such as contour shape, internal edges or corners – thus allowing objects to be identified based on their features.

In this paper, we are concerned with finding a target in real time for a vision based guidance system. If a target is approached at constant speed, the size of the projected image of the target increases as a negative linear inverse, which is faster than exponential. In other words, the perceived size of the target from the perspective of the camera will start out very small and then increase super-exponentially as it is approached. This is illustrated geometrically in Fig. 1.

• *S. Heinrich is with the Department of Computer Science, North Carolina State University, Raleigh, NC, 27606.*
E-mail: sbheinri@ncsu.edu

It is our goal to design a system for recognizing targets while they are still quite small, thereby allowing adequate time for a vision based guidance system to adjust its flight path towards a mobile target, which might have moved from its location at the time of launch.

Specifically, we propose to do this by making use of a previous survey image containing the target and its surroundings. Given a bound on the mobile target's velocity, this defines a search range from the position in the survey image, which we utilize to restrict our search in the image taken from the missile as it approaches.

2 Method

From a sufficient elevation, most areas of the ground can be well approximated by a ground plane. A ground plane induces a homography between the survey image and the search image, which is a global function that transfers visible points in one view to the other under perspective projection regardless of imaging viewpoints [HZ03] [Bro92].

The first step in our algorithm is to compute an estimate of the homography (also known as a projectivity [Bro92] [ZF03]) because it is a quick and reliable way to register aerial images taken from different viewpoints under perspective effects. We do not expect the homography to be perfect because the background won't be truly planar, but it works as a good estimate in most cases.

After finding a homography, the process of transferring the search image into the view of the survey image is simple matrix multiplication and interpolation. We expect that the target can be found in the search image somewhere nearby to its previously observed location.

Because the target will have moved parallel to the ground plane (if at all), and the registration accounts for perspective and scale changes, the target's image in the registered search image can be related to the target image in the survey image by a simple translation and rotation (assuming the ground is roughly planar).

However, at the small image sizes we are expecting, the target image is dominated more by rasterization artifacts

and shadows than features which are actually useful for classification, so that even if the correct rotation and translation were known, a direct template match would not work well (see Fig. 6).

One feature that can still be reliably extracted from a small target such as this is its histogram, which is conveniently invariant to rotation. This motivates the design of our histogram shift search for locating the target based on its histogram.

Our algorithm can now be summarized as consisting of the following four steps:

1) **Find candidate correspondences between views.** A set of correspondences are found between salient points in both views found by comparing the local brightness in a basis with unified scale and orientation with a feature matcher that is robust to small errors.
2) **Estimate the homography between the two views.** A solution is achieved by non-linear least squares and RANSAC [FB87], using the epipolar constraint on the found correspondences.
3) **Transfer the search image to the survey view.** After finding the homography, the search image can be directly transfered (aka, registered) into the same coordinate system of the survey image.
4) **Locate the target by histogram shift search.** The target is located in the registered search image by searching nearby to the previously known location. It is identified by a similarly scaled region having similar histogram.

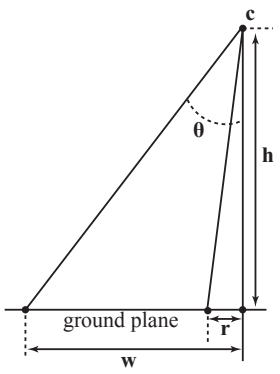

Fig. 1. **Relative size of target image.** A camera with focal point c looking directly at a target having radius r on the ground plane. If the camera field of view is 2θ, and the height of the camera from the ground plane is h, then the visible width of the ground is $2w = 2h\tan\theta$. The relative image size of the target is $r/(h\tan\theta)$.

2.1 Finding Correspondences

In order to find correspondences, we first find a set of salient points in the images and then match them based on local neighborhoods. Each salient point has an associated scale and rotation (in the image plane). For our salient points, we use the set of multi-scale Harris corners [HS88] [MS04] where the *characteristic scale* [Low04] is close to the scale used to compute the local structure tensor of the multi-scale corners. This prevents corner points from generating a large number of salient points across scale-space.

For the dominant orientation, we use the peak of the kernel density estimate of the local gradient orientation in the characteristic scale as the dominant orientation. This has a simple implementation achieved by accumulating the Gaussian weighted local image gradient angles into a buffer, performing a 1D Gaussian blur on that buffer, and taking the angle corresponding to the bin with the largest value.

Because we only use the dominant orientation for finding a consistent basis to do matching, it is not necessary for the found angle to agree with the human perception of the visually dominant orientation; as long as it is consistent for the corresponding points, it will suffice.

In order to find correspondences from the set of salient points, we use a grid-based feature matcher that is robust to small errors in the estimated scale and orientation to compare each salient point with salient points in the other image of similar scale. Finally, we use the set of cross-validated matches as correspondences. Further details on our approach to finding correspondences can be found in [KHS+08a], [KHS+08b] and [HKS+08].

2.2 Estimating the Homography

Discovering the homography relating two views is similar to computing the epipolar geometry, and can also be represented by an invertible 3×3 matrix [HZ03]. Denoting this matrix $\mathbf{H}$, points are transferred according to the following homogeneous equation,

$$(x : y : 1)\,\mathbf{H} = (x' : y' : w')^{\mathsf{T}} \tag{2.1}$$

where (x, y) and $(x'/w', y'/w')$ are the image space coordinates of the same world point imaged in two views.

Because it has only 8 degrees of freedom [Hec89] [HZ03], there is a unique solution given 4 corresponding points which can be computed by back substitution [Hec89].

Given more than 4 correspondences, a least squares solution can be found using the Direct Linear Transform (DLT), or iterative methods such as the Gauss-Newton method or the Levenberg-Marquardt method [HZ03].

Finding correspondences can be error prone, especially when the images may contain scale changes, viewpoint/perspective changes, moving features, lighting changes, specular lighting changes induced by viewpoint changes, and 3D features that violate the assumption of background planarity.

Therefore, we do not rely on a simple least squares estimate based on all correspondences. Instead, we use RANSAC [FB87] to find the least squared estimate of the largest random sample consensus, which we improve using the Gauss-Newton on the resulting inlier set. This method is similar with that suggested in [HZ03].

2.3 Histogram Shift Search

We present the histogram shift search for finding a target with known scale and unknown rotation given a rough guess of the target's position. This method is also robust to other common forms of geometric distortion, such as affine, perspective, and lens distortion, as long as they do not significantly alter the histogram or overall scale.

The basic algorithm is simple: given some known scale and histogram of a target, and a very rough guess of the target's location, generate random seed locations distributed around the initial guess. For each random seed, use the *histogram shift procedure* to shift the window center until the difference between the local window histogram and the target histogram is locally minimized. Then return the best overall minimum.

This search method is particularly well suited to recognizing small objects when there are too many degrees of freedom for geometric distortions to efficiently enumerate, because the histogram of the image of an object's projection is invariant to rotations in the image plane, very robust to rotations out of the plane, and efficient to calculate for small regions.

2.3.1 Histogram Shift Procedure

The histogram shift procedure is an iterative method of gradient descent for shifting an initial guess of window position towards the position that locally minimizes the difference between the histogram of the local window and a reference histogram.

Given a circular window centered at $\mathbf{x} = (x, y)$, we define the triple-band histogram of the circular region in an image with known radius as $\mathcal{H}(\mathbf{x})$. As a similarity metric between histograms, we use the L^1-norm across all bins in all color bands. If we denote the reference histogram as $\mathcal{H}_{ref}$, then our objective is to find the $\mathbf{x}$ that minimizes f in

$$f(\mathbf{x}) = ||\mathcal{H}(\mathbf{x}) - \mathcal{H}_{ref}||_1 \tag{2.2}$$

Because there is no closed form for partial derivatives of the objective function, we use the backward finite difference approximation of the total derivative on the image grid as a means for approaching the minimum. The approximated partial derivatives are given by

$$\frac{\partial f}{\partial x} \approx f(x, y) - f(x-1, y) = \Delta f_x \tag{2.3}$$

$$\frac{\partial f}{\partial y} \approx f(x, y) - f(x, y-1) = \Delta f_y \tag{2.4}$$

Because these are only first order approximations and the state space may be highly noisy (see Fig. 7 for an example), we do not generally trust that the values will be accurate, so we use only the sign of the approximated derivatives and take individual pixel steps towards the minimum. This also reduces computational complexity by avoiding the need for image interpolation.

Given the position (x^j, y^j) in iteration j, the algorithm for updating x^j is defined by the following recurrence (and similarly for y^j),

$$x^{j+1} = \begin{cases} x^j + 1, & \Delta f_x < 0 \text{ and } x^{j-1} \neq x^j + 1; \\ x^j - 1, & \Delta f_x > 0 \text{ and } x^{j-1} \neq x^j - 1; \\ x^j, & \text{o/w.} \end{cases} \tag{2.5}$$

Because this procedure is not guaranteed to converge, we impose a cutoff on the number of iterations to prevent infinite loops. One might also choose to terminate if the same location is ever revisited.

3 Results

We have created a scene to test our algorithm by taking an aerial photo of an urban environment and extruding 3D buildings. We have also added some 3D vegetation that is simulated by opacity mapped noisy geospheres. For our target we have juxtaposed an appropriately scaled model of an M1 Abrams tank. Rendering was done using Mental Ray 3.5 with physically simulated sky, atmospherics, and shadows.

We have generated two views which we denote the survey view and search view (see Fig. 2). The tank's position, orientation, and turret rotation change between views. The location of the sun changes, as does the camera view location. The survey view is roughly downward-facing, but not perfectly.

We found 83 correspondences using our algorithm, 73 of which fit with the found homography solution with less than 1 pixel error, 7 of them have approximately 2 pixels or less of error, 1 of them had about 8 pixel error, 1 had 53 pixel error, and one had 72 pixel error.

Another popular method for finding correspondences is SIFT (Scale-Invariant Feature Transform) [Low04], and we compare our method of finding correspondences on this image pair vs SIFT using a graph of correct correspondences found vs. precision in Fig. 3. Note that we do not use recall because the salient points found by these two algorithms differs. The implementation of SIFT that we use is SIFT++ (freely available online at http://vision.ucla.edu/~vedaldi/code/siftpp/siftpp.html).

Precision is the percent of correct correspondences returned by the algorithm. A correspondence is considered correct if the difference between the corresponding point and the projected point using the ground truth homography is less than 0.02 (in normalized coordinates). The ground truth homography was computed by taking the least-squares homography from user-selected correspondences belonging to the ground plane.

These results indicate that on this test image, SIFT was not able to surpass 25% precision, making it unusable for recovering the true homography. In contrast, using our method we were able to obtain sufficient correspondences at greater than 90% precision.

From the correspondences, we generated a robust estimate of the homography using Gauss-Newton non-linear least squares regression and RANSAC [**?**]. For the RANSAC parameters, we used 10% of the correspondences for making an initial pick, 40% of the correspondences for the minimum number in an acceptable solution, a tolerance of 0.02 for inliers (using normalized coordinates), 500 iterations of RANSAC, and 20 iterations of Gauss-Newton non-linear least squares regression.

The center and radius of the tank in the survey image was manually chosen by the operator (Fig. 4). Then the search image was transferred into the view of the survey image, as shown in Fig. 5 (it is shown here being overlapped on top of the survey image to illustrate the correctness of the homography). Note that the 3D building tops are not expected to align perfectly because they do not belong to the ground plane.

(a) (b)

Fig. 2. **Test images from simulated urban environment.** (a) Survey image. (b) Search image.

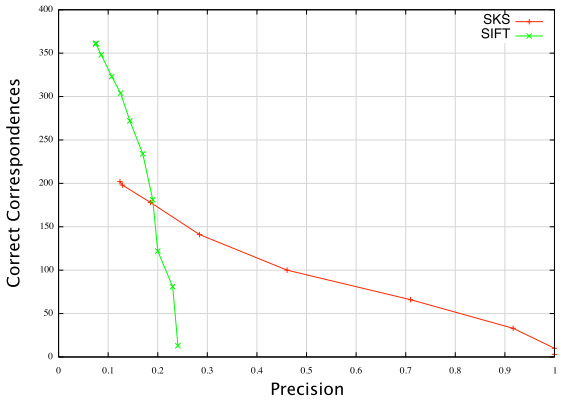

Fig. 3. **Correct correspondences vs precision.** Number of correct correspondences found by the algorithm vs. precision. Note that the total number of correspondences found is correct correspondences divided by precision. This graph shows that SIFT never finds more than 25% correct correspondences, whereas our method (SKS) can generate a usable number of correspondences all the way up to 90% precision.

(a)

Fig. 4. **Selected target.** The target to search for is selected manually from the survey image by an operator, shown by the red circle.

The histogram search method was used to locate the target in the registered search image, which has approximate size of 16×7 pixels (see Fig. 6). The target found by this procedure is indicated by the red circle in Fig. 5.

To represent the histograms, we used 256-bin histograms for each color channel in sRGB space. For the search parameters, we generated 50 random seeds with a ± 40 pixel error of the expected location. To prevent infinite loops, we used a maximum cutoff of 15 iterations in the histogram shift procedure.

(a)

Fig. 5. **Search image registered to survey view.** Using the found homography, the search image is registered to the survey view. Green trails indicate the paths taken using the histogram shift procedure from each seed point, and the final result of the search is shown by the red circle.

We found it useful to visualize the function $f(\mathbf{x})$ across the state space of the registered search image. This is shown in Fig. 7, which makes it clearly visible that there is one unique global minimum in the image having the most similar histogram to the original target.

Finally, the homography is inverted and used to transfer the found target position back into the original search image, as shown in Fig. 8.

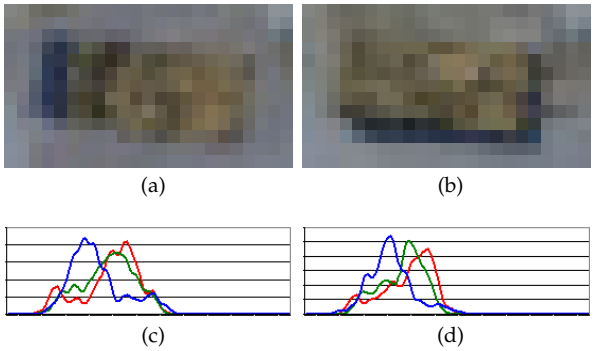

Fig. 6. **Closeup views of tank image.** (a) region containing tank in survey image, after being rotated and resampled. (b) region containing tank in registered search image. The dimensions of the tank are roughly 16×7 pixels. At this small scale, internal features are dominated by shadows and sampling artifacts, so that even a perfectly matched correlation would not work well. However, the histograms (shown below) are still similar.

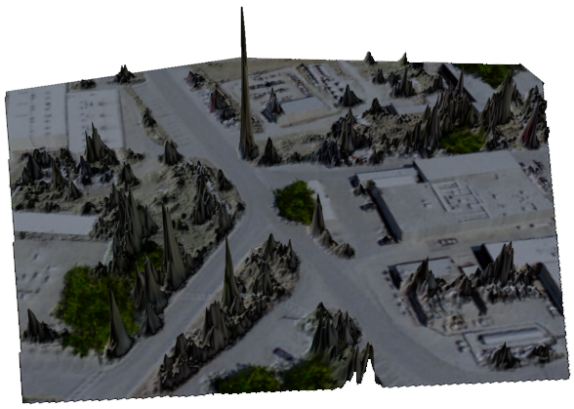

Fig. 7. **Histogram difference state space.** The relative value of $f(\mathbf{x})$ is indicated by the height of the displacement at each pixel location in the registered search image. The surface is colored by the image so that the correspondence between state space and image is readily apparent. In this visualization, the function is actually inverted so that the lowest values appear as the highest peaks. There is clearly one maximal peak corresponding to the true target location.

4 Conclusions

For missiles intended for mobile targets, automatic target detection from visual queues could be used to increase missile precision. However, because the projected image size of a target grows faster than exponentially as the target is approached linearly, being able to do detection on a *slightly* smaller target can be used to detect the target from a *greatly* larger distance.

A missile's flight path cannot be corrected after it is too close, so the minimal size that reliable recognition can be achieved for this application is critical.

Therefore, as a means for making longer-range detection possible, we have presented a method for detecting targets with very small projected images that is fast and robust. We have demonstrated the ability of our algorithm to function under the effects of perspective distortion, with 3D clutter, and under changed lighting conditions. Further, we did not assume a static target, allowing the tank target to move, change orientation, and rotate the turret.

In our test example, a target having size of approximately 16×7 pixels was successfully located, which according to Johnson's Criteria [Joh85], is not quite large enough even for a human (which are generally far more capable of target identification than computer visions systems) to be capable of identifying with 50% confidence.

Fig. 8. **Target identified in search image.** The result of automatic target recognition, shown as the red ellipse.

A limitation of our method is the assumption that the background is mostly planar in the computation of a homography. In most cases this assumption is reasonable, but not always. Because our primary use of the homography is for choosing an initial guess for the histogram search, it does not need to be overly accurate.

However, we also use the homography to resample the search image, so if the assumption of planarity is *wildly* inaccurate, then the homography estimation might be so terrible that the guessed location be completely off, or the scale of the target in the registered image was drastically changed, causing the histogram search method to fail. This might occur, for example, if the target was located on a steep mountain and both views contained overlapping parts of the mountain foreground and foothills and/or sky.

REFERENCES

[BMP02] Serge J. Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. Technical report, Berkeley, CA, USA, 2002.

[Bro92] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM Comput. Surv.*, 24(4):325–376, 1992.

[FB87] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. pages 726–740, 1987.

[Hec89] Paul S. Heckbert. Fundamentals of texture mapping and image warping. Master's thesis, Berkeley, CA, USA, 1989.

[HKS+08] Stuart B. Heinrich, Karthik Krish, Wesley Snyder, Siamak Khorram, and Halil Cakir. Matching of mobile targets in overlapping aerial images. In *Proc. ASPRS*, 2008.

[HS88] C. Harris and M. Stephens. A combined corner and edge detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[HZ03] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge U. Press, second edition, 2003.

[Joh85] J. Johnson. Analysis of Image Forming Systems. In R. B. Johnson and W. L. Wolfe, editors, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 513 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 761–+, 1985.

[KHS+08a] Karthik Krish, Stuart B. Heinrich, Wesley Snyder, Halil Cakir, and Siamak Khorram. Fast global registration of overlapping images using accumulative image features. Technical report, Center for Earth Observation, North Carolina State University, 2008. TR-225.

[KHS+08b] Karthik Krish, Stuart B. Heinrich, Wesley Snyder, Halil Cakir, and Siamak Khorram. A new feature based image registration algorithm. Technical report, Center for Earth Observation, North Carolina State University, 2008. TR-224.

[Low99] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2*, page 1150, Washington, DC, USA, 1999. IEEE Computer Society.

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[MS04] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, 2004.

[Pop94] Arthur R. Pope. Model-based object recognition - a survey of recent research. Technical report, Vancouver, BC, Canada, Canada, 1994.

[ZF03] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21:977–1000, 2003.