THE JOURNAL OF NAVIGATION (2016), **69**, 1215–1233. © The Royal Institute of Navigation 2016
doi:10.1017/S0373463316000187

Multi-Region Scene Matching Based Localisation for Autonomous Vision Navigation of UAVs

Zhenlu Jin[1,2], Xuezhi Wang[3], Bill Moran[3], Quan Pan[1] and Chunhui Zhao[1]

[1](*School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi, China*)
[2](*Department of Electrical and Electronic Engineering, University of Melbourne, Australia*)
[3](*School of Engineering, RMIT University, Australia*)
(E-mail: zhaochunhuinpu@126.com)

A multi-region scene matching-based localisation system for automated navigation of Unmanned Aerial Vehicles (UAV) is proposed. This system may serve as a backup navigation error correction system to support autonomous navigation in the absence of a global positioning system such as a Global Navigation Satellite System. Conceptually, the system computes the location of the UAV by comparing the sensed images taken by an on board optical camera with a library of pre-recorded geo-referenced images. Several challenging issues in building such a system are addressed, including the colour variability problem and elimination of time-varying details from the pairs of images. The overall algorithm is an iterative process involving four sub-processes: firstly, exact histogram matching is applied to sensed images to overcome the colour variability issues; secondly, regions are automatically extracted from the sensed image where landmarks are detected via their colour histograms; thirdly, these regions are matched against the library, while eliminating inconsistent regions between underlying image pairs in the registration process; and finally the location of the UAV is computed using an optimisation procedure which minimises the localisation error using affine transformations. Experimental results demonstrate the proposed system in terms of accuracy, robustness and computational efficiency.

KEYWORDS

1. Scene Matching. 2. Colour Constancy. 3. Landmark Detection. 4. Multi-Region.

Submitted: 28 July 2015. Accepted: 14 March 2016. First published online: 18 April 2016.

1. INTRODUCTION.

Navigation technology plays a crucial role in the deployment of Unmanned Aerial Vehicles (UAV). A conventional Inertial Navigation System (INS) may accumulate large errors over time and Navigation Error Correction (NEC) is a necessary procedure to maintain the navigation error within acceptable bounds. While NEC can be obtained via the Global Position System (GPS) or other Global Navigation Satellite System (GNSS), a vision-based navigation system may serve as an alternative for autonomous navigation of UAVs in the absence of GPS (Wang et al., 2013).

Approaches for localising a UAV using a vision-based navigation system are usually divided into three categories (Bonin-Font et al., 2008): mapless localisation (such as visual odometry (Williams and Reid., 2010)), map-building-based localisation (such as Simultaneous Localisation And Mapping (SLAM) (Nemra and Aouf., 2009)), and map-based localisation (such as scene matching methods (Lincheng et al., 2010)). These three kinds of vision navigation methods have different advantages, drawbacks and areas of applicability. On the one hand, the mapless and map-building-based methods just need a camera mounted on the UAV, while the estimation errors of inter-frame motions would accumulate severely (Williams and Reid., 2010). On the other hand, scene matching methods require an extra library of pre-recorded geo-referenced images, but allow absolute locations of UAVs to be obtained (Li et al., 2009) without accumulative errors.

A vision-based NEC system, as illustrated in Figure 1, is our final goal to improve navigation precision by combining localisation results of scene matching system, INS, and GPS, with the focus on scene matching, where I'_k and I_k denote the sensed and reference image, respectively. k signifies time index. P'_k represents the calculated position of I'_k in I_k, and P^0_k is the prior location of the UAV. Geo-referenced images, pre-recorded from areas of interest, are assumed to be available. An essential part of the scene matching system is image matching, and this has been explored in many papers, including correlation methods (Zhao et al., 2006), edge methods (Ling et al., 2009) and point correspondences methods (Bay et al., 2008).

In order to establish a practically robust vision-based NEC system, several issues regarding UAV localisation via image scene matching need to be addressed. Firstly, the pair of images to be registered are taken in different conditions at different times. As time-varying objects may be involved in the scene, registration may yield large errors if standard image features are used. Secondly, without specific knowledge of the time invariant objects in the scene, a statistical model for selecting time invariant objects is needed. Thirdly, a statistical feature in the scene of underlying images, which is used to geometrically register image pairs containing time-varying objects, is preferable for an efficient autonomous navigation scheme. Clearly, these research challenges need to be adequately addressed.

As the position of the sun varies through the day and over seasons, the colours of sensed images taken at different times may appear quite different from those of the reference images. This colour variability can significantly influence the precision and reliability of scene matching. Consequently, colour constancy processing is very important in mitigating colour variability issues (Van De Weijer et al., 2007). In Agarwal et al. (2006), the state of the art colour constancy algorithms are divided into two categories: pre-calibrated and data-driven approaches. The latter categories include methods such as grey world assumptions (Buchsbaum, 1980), Retinex-based white patch approaches (Barnard et al., 2002), grey edge algorithm (Van De Weijer et al., 2007), machine learning methods (Ebner, 2004), etc. When the contents in two images are similar, colour constancy methods with satisfactory performances can be found in Morovic et al. (2002) and Jin et al. (2015).

For many applications, partitioning an image into multi-regions rather than using the entire image for scene matching between a pair of images significantly reduces computational complexity and improves localisation precision (Calloway et al., 1990; Li

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

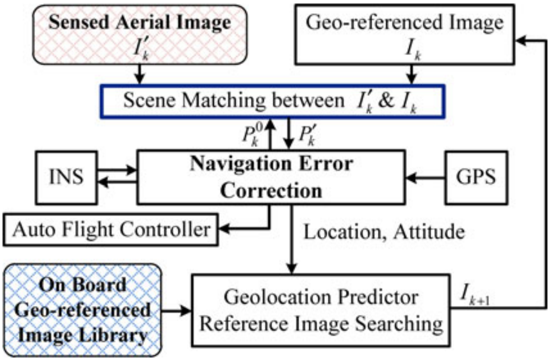


Figure 1. Illustration of a Vision based NEC System.

et al., 2009). For visual NEC applications, the multi-region scene matching approach also appears as a necessary technique to remove local dissimilarity between a pair of images from consideration, though this remains a difficult problem. A novel method to extract multi-regions is to select salient regions by mimicking the visual attention mechanism of primates (Siagian and Itti, 2009). The work in Jin et al. (2013) describes a multi-feature fusion visual saliency model that incorporates image features suitable for scene matching. Multi-region extraction can also be accomplished by selecting landmarks, which are reasonably assumed to be time-invariant for scene matching. With the colour histogram as the only extracted feature of landmarks, preliminary work on this topic is reported in Jin et al. (2014b).

On the other hand, multi-region scene matching may yield large registration errors or limited numbers of successfully matched sub-regions as less information is used than matching the entire image. In Lo and Gerson (1979), a least-squares estimator is adopted to estimate the affine transformation so as to perform positioning by global optimisation. To achieve an optimal localisation result, geometric constraints on multi-regions are imposed to fix reference points and infer the most accurate location (Li et al., 2009), which improves robustness to shifts between sensed images and reference images. Following Li et al. (2009), an efficient method for sub-region selection and matching error reduction is proposed in Jin et al. (2014a) based on visual saliency computation and affine constraints; this performs well even under errors in scaling and rotation between sensed images and reference images.

A multi-region scene matching-based UAV localisation algorithm is proposed for vision navigation systems, which combines the aforementioned three image processing techniques. The contributions of this paper are:

- The approach is fairly general in the sense that we use a statistical feature for image registration, which enables automated scene matching using a generic landmark selection method and tolerates different non-landmark objects in the pair of images.
- A recursive colour constancy algorithm that exploits the location, scaling and rotation parameters extracted from the NEC system is presented to adaptively reduce the colour variation between an image pair to be matched. This method

is incorporated with the statistical landmark selection using the colour histogram of landmarks and a multi-region scene matching algorithm.

- An estimation procedure that determines the location of the UAV using a set of registered sub-region centres with errors using affine transformations is presented. This approach is optimal in the sense that it minimises estimation errors by taking all possible combination of registered region locations into account. As demonstrated by experimental results, the estimation procedure is robust to shifting, rotation and zooming between image pairs in image registration.

In principle, any image matching method, including the widely used feature point based methods, can be used to register the selected landmark regions. Landmark selection effectively eliminates the regions possibly containing time-varying objects from consideration before image matching. Clearly, the multi-region scene matching step has a reduced number of feature points and computational load compared with the registration using the entire image.

2. THE PROPOSED VISION-BASED LOCALISATION SYSTEM.

The overall idea of a reliable multi-region automatic scene matching system is illustrated in Figure 2, where the NEC system provides the prior location of UAV P_k^0, which may be derived from the last localisation result P'_{k-1} or the result of INS, GPS, or a combination of these (Jwo et al., 2013).

The first phase addresses image pre-processing and colour constancy processing. Image pre-processing reduces the scaling, rotation, and perspective differences between the sensed image I'_k and the reference image I_k. The colour constancy processing adjusts the colour of a sensed image to overcome colour variability problems with the reference image. The Exact Histogram Matching (EHM) algorithm in Morovic et al. (2002) achieves colour constancy if the scene of the sensed image can be found exactly from the geo-referenced image. The UAV location uncertainty arising from erroneous scene matching degrades the performance of colour constancy processing, which in turn degrades the scene matching and thus the localisation performance. Therefore, we adopt an iterative local EHM procedure. The sensed image is partitioned into a set of non-overlapping windows. The corresponding sub-images are extracted from the reference image with shift, scaling and rotation compensation as shown in Figure 2. This process may iteratively update the calculated location P_k, scaling b_k and rotation α_k parameters of the sensed image with respect to the reference image before they are used for navigation.

The second phase is to extract sub-regions from the sensed image after colour constancy processing $I'_{\mathrm{R},k}$. Sports fields, buildings, roads and rivers are treated as “time-invariant” landmarks. As the shape and orientation of a landmark is generally not known, we use colour histograms as a generic feature to distinguish landmarks from other objects. The colour histogram of each type of landmark, denoted by $H_{\mathbf{I}_{\text{Training}}}$, can be obtained by training with a landmark pattern library. The training library is selected manually from the geo-referenced image library. A likelihood map is calculated for each kind of landmark from the sensed image based on how its colour intensity matches that of the training data as preliminarily discussed in Jin et al. (2014b). In this paper, we extend previous work to include the automatic sub-region extraction from the likelihood maps. The candidate sub-regions containing landmarks $\{I'_{1,k}, I'_{2,k}, \ldots, I'_{M,k}\}$ are detected automatically by applying experimentally

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

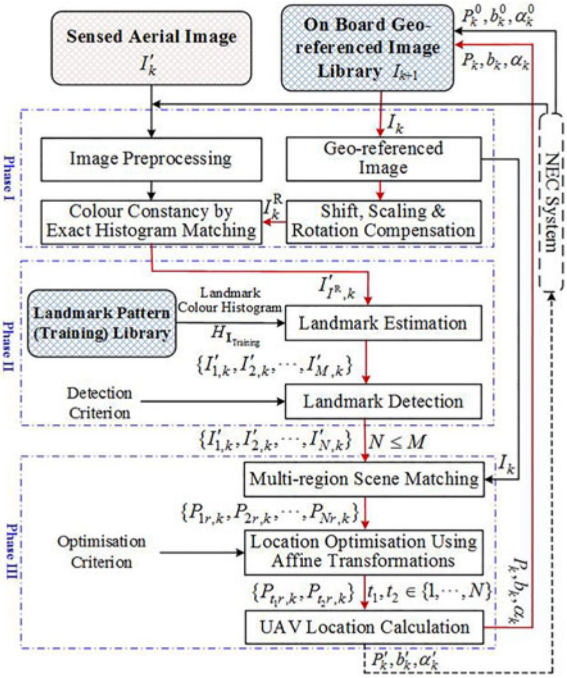

Sensed Aerial Image
I'_k
On Board Geo-referenced Image Library I_{k-1}
P_k^0, b_k^0, α_k^0
P_k, b_k, α_k
I_k
Phase I
Image Preprocessing
Geo-referenced Image
Colour Constancy by Exact Histogram Matching
I_k^R
Shift, Scaling & Rotation Compensation
NEC System
Phase II
$I'_{I^R,k}$
Landmark Pattern (Training) Library
Landmark Colour Histogram
$H_{I_{Training}}$
Landmark Estimation
$\{I'_{1,k}, I'_{2,k}, \cdots, I'_{M,k}\}$
Detection Criterion
Landmark Detection
$\{I'_{1,k}, I'_{2,k}, \cdots, I'_{N,k}\}$
$N \le M$
Phase III
I_k
Multi-region Scene Matching
$\{P_{1r,k}, P_{2r,k}, \cdots, P_{Nr,k}\}$
Optimisation Criterion
Location Optimisation Using Affine Transformations
$\{P_{t_1r,k}, P_{t_2r,k}\}$
$t_1, t_2 \in \{1, \cdots, N\}$
P_k, b_k, α_k
UAV Location Calculation
P'_k, b'_k, α'_k

Figure 2. The Idea of Multi-region Scene Matching.

predetermined thresholds. Eventually $N \le M$ sub-regions are selected $\{I'_{1,k}, I'_{2,k}, \ldots, I'_{N,k}\}$, where multiple (partially) overlapped regions containing different type of landmark are combined into one region with the centre location given by the geometrical average over all overlapped region centres. For example, a river region might overlap a road region. Colour constancy processing improves the quality of landmark estimation.

The final phase consists of a multi-region scene matching process and a UAV localisation optimisation procedure using affine transformations. The former is implemented in parallel so that the locations of sub-regions $\{P_{1r,k}, P_{2r,k}, \ldots, P_{Nr,k}\}$ are obtained simultaneously. Because of the pre-processing, registration of the sensed and reference images under identical view angles is possible, and the affine constraints are preserved for any three matched sub-regions. As the registered locations are actually inaccurate, three regions $\{P_{t_1r,k}, P_{t_2r,k}, \ldots, P_{t_3r,k}\}$ are chosen from all the sub-regions by an optimisation criterion, for instance, the minimum relative estimation error. These three sub-regions form reference points from which the optimal location P_k is calculated using affine transformations. Initial work on this idea was reported in Jin et al. (2014a). Note that based on the set of matched sub-region centres, we can also localise the UAV in the world coordinate system.

The library of geo-referenced images should be built to enable a fast image query process. For example, multiple level images for a single zone may be stored as a function of altitude and ground resolution and images between adjacent zones are required to be partially overlapped. Apart from geo-coordinates of centre points and boundary points, library images could also be indexed by features (Sim et al., 2002), such as colour distribution, texture, feature points, edge map, etc.

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

On receiving a sensed image, searching for the corresponding geo-referenced image from the library can be done in one of two ways. Firstly, if we have prediction or prior knowledge of the current location of UAV, a Bayesian estimator (Conte and Doherty, 2011) may be used to search the geo-referenced image locally from its prior geo-coordinates. In the rare case that prior information is not available, an exhaustive (feature-based) search over the entire database could be required to initialise the geo-coordinate tracker, which will work recursively as more UAV location estimates are obtained. Further research on this topic will be reported elsewhere.

3. AUTOMATED MULTI-REGION SCENE MATCHING METHOD.

The procedure of the proposed method (see Figure 2) consists of the following steps:

Step 1 Pre-process the sensed image with the location and attitude information of the UAV provided by the NEC system.

Step 2 Apply local EHM-based colour constancy processing on the sensed image according to the corresponding reference image with shift, scaling, and rotation compensation.

Step 3 Extract sub-regions from the sensed image by selecting landmarks.

Step 4 Conduct multi-region scene matching if the number of landmarks is greater than three. Otherwise, proceed with the scene matching using the entire sensed image and jump to Step 8.

Step 5 Compute the locations of sub-regions using affine transformations.

Step 6 Select three sub-regions as reference points and infer the UAV location.

Step 7 Iterate Step 2–7 until the UAV location converges.

Step 8 Output the location of the UAV to the NEC system.

Let the estimated UAV locations in two consecutive iterations (n and n–1) be denoted by P_k^n and P_k^{n-1}. The stopping criterion for convergence of the localisation iteration using an aerial image is given by

$$||P_k^n - P_k^{n-1}|| < \varepsilon \tag{1}$$

where $||\cdot||$ signifies the Euclidian distance and $\varepsilon > 0$ is a small number.

We denote the proposed algorithm by MRSM-CC (Multi-Region Scene Matching with Colour Constancy). It operates automatically without human intervention.

3.1. *Colour Constancy Processing on Sensed Image.*

The EHM method (Morovic et al., 2002) transfers the original histogram of an original grey image to the target histogram of a target grey image as illustrated in Figure 3.

The EHM method is applied locally and iteratively on the sensed image to achieve colour constancy separately in the R, G and B channels as illustrated in Figure 4, where the sensed image I' is the "original image" and the geo-referenced image I^R is the "target image". I^R is obtained from the reference image I with the UAV location parameters provided initially by NEC and in the subsequent iterations by the localisation outcomes.

In our experiments, the sensed image I' is divided into a set of 5 × 5 non-overlapping windows $\{I'_1, I'_2, \ldots, I'_{25}\}$. Their corresponding sub-images $\{I_1^R, I_2^R, \ldots, I_{25}^R\}$ are also extracted from the reference image I^R. In general, the sizes of I' and I^R can be different.

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

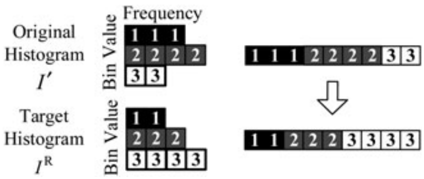


Figure 3. Illustration of the EHM algorithm.

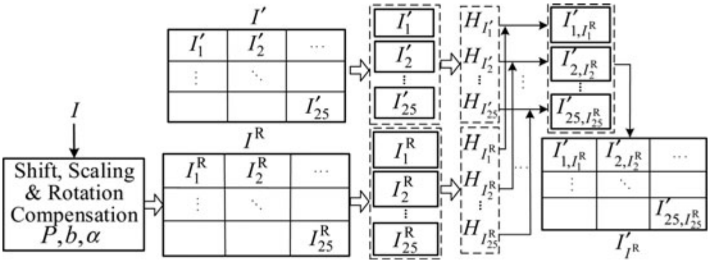


Figure 4. Illustration of the local EHM operation.

If this is the case, differences between the pair of images such as shifting, scaling, etc. should be considered for the determination of I^{R}. In addition, the histograms $\{H_{I_1^{\mathrm{R}}}, H_{I_2^{\mathrm{R}}}, \ldots, H_{I_{25}^{\mathrm{R}}}\}$ of the sub-images from the reference image are normalised by

$$H_{I_i^{\mathrm{R}}}^{\mathrm{N}} = \frac{\sum H_{I_i'}}{\sum H_{I_i^{\mathrm{R}}}} H_{I_i^{\mathrm{R}}}, \; i = 1, \ldots, 25. \tag{2}$$

Look-Up Tables (LUT) are calculated to transfer $\{H_{I_1'}, H_{I_2'}, \ldots, H_{I_{25}'}\}$ to $\{H_{I_1^{\mathrm{R}}}^{\mathrm{N}}, H_{I_2^{\mathrm{R}}}^{\mathrm{N}}, \ldots, H_{I_{25}^{\mathrm{R}}}^{\mathrm{N}}\}$. The transformation is performed one by one along with histogram bins in ascending order. At each step, two variables are calculated: the number of pixels N_{req} (m, n) required to be assigned the value n, and the number of pixels N_{rem} (m, n) remaining unchanged as the value m. The number of pixels to be assigned a new value at row m and column n in an LUT is $\min(N_{\mathrm{rem}}\ (m, n), N_{\mathrm{req}}\ (m, n))$. For example, as shown in Figure 3 at bin number $m = 1$ for I' and $n = 1$ for I^{R}, we have $N_{\mathrm{req}}\ (m, n) = 2$ and $N_{\mathrm{rem}}\ (m, n) = 3$. Therefore, the number of pixels to be assigned to the bin $n = 1$ is $\min(N_{\mathrm{rem}}\ (m, n), N_{\mathrm{req}}\ (m, n)) = 2$. The interested reader is referred to (Morovic et al., 2002) for more detail. According to the LUTs, $\{I_1', I_2', \ldots, I_{25}'\}$ in the sensed image are assigned new colour distributions. The sensed image after colour constancy processing $I'_{I^{\mathrm{R}}}$ is obtained by combining $\left\{I'_{1,I_1^{\mathrm{R}}}, I'_{2,I_2^{\mathrm{R}}}, \ldots, I'_{25,I_{25}^{\mathrm{R}}}\right\}$.

The more accurate the location and attitude of the UAV, the better the performance of this local EHM procedure is. After landmark detection and multi-region scene matching, the local EHM process is repeated on the sensed image, and this process is iterated until the location of the UAV converges.

3.2. *Automatic Sub-Region Extraction by Landmark Detection.* The colour histogram of each type of landmark (roads, sport fields, buildings and rivers) is obtained

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

from a collection of training images. A landmark detection method based on likelihood thresholding is then applied to extract the sub-regions of this type of landmark from the sensed image for scene matching.

Let B denote the event that a pixel belongs to a type of landmark. The colour histogram of the training set $\mathbf{I}_{\text{Training}}$ for this type of landmark is denoted by $H_{\mathbf{I}_{\text{Training}}}$. Its probability density function is approximated by normalising the colour histogram and denoted by p_B.

The colour intensity at the (i, j)th pixel in the sensed image is represented by $I'_{i,j} = [r, g, b]^{\mathrm{T}}_{i,j}$. The likelihood that $I'_{i,j}$ originates from the landmark intensity distribution B is given by the conditional probability density function $p(I'_{i,j}|B)$ which is approximated by the normalised colour histogram of landmark $p_B\,(I'_{i,j})$. Using Bayes' rule, it is straightforward to conclude that $p(B|I'_{i,j})$ for the (i,j)th pixel on the test image is proportional to the likelihood of the presence of a type of landmark ($P(B)$ >0),

$$P(B|I'_{i,j}) \propto p_B(I'_{i,j})P(B) \tag{3}$$

Therefore, we are able to calculate a likelihood map of a given landmark based on the sensed image.

While the feature of a landmark is characterised by the colour histogram, its spatial information is discarded. Taking spatial information into account will certainly lead to a better landmark characterisation. For instance, if the colour of a pixel representing a building is white, the colours of its neighbouring pixels are also expected to be white.

Let $I'_r\,(i, j)$ be the average colour intensity in the neighbourhood $\mathcal{N}_{i,j}$ of the (i,j)th pixel, i.e.,

$$I'_r(i,j) = \frac{1}{|N_{i,j}|} \sum_{s,h \in N_{i,j}, s,h \neq i,j} I'(s,h) \tag{4}$$

where $|N_{i,\,j}|$ denotes the cardinality of $N_{i,\,j}$. This quantity is used to represent the local spatial colour feature of the underlying pixel.

Let $\bar{p}_B(I'_r(i,j))$ be the normalised colour histogram on a type of landmark at the (i, j) th pixel, which can also be computed from the landmark training set. In a similar way to Equation (3), the probability of the event B given $I'_r\,(i, j)$ is proportional to the colour density in the neighbourhood of the (i, j)th pixel; that is,

$$P(B|I'_r(i,j)) \propto \bar{p}_B(I'_r(i,j)). \tag{5}$$

Considering Equations (3), (4) and (5), the probability that the underlying pixel belongs to a type of landmark (i.e., event B) conditioned on the two independent events is given by

$$P(B|I'_{i,j}, I'_r(i,j)) \propto P(B|I'_{i,j})P(B|I'_r(i,j)) \propto p_B(I'_{i,j})\bar{p}_B(I'_r(i,j)). \tag{6}$$

Using Equation (6), a likelihood map can be calculated for each type of landmark from the sensed image. Therefore, for a given probability threshold, those sub-regions containing landmarks can be extracted from the computed probability maps.

3.3. *Location Optimisation Using Affine Transformations.* Various image registration techniques available in the literature can be applied for multi-region scene matching. In this work, we use the Normalised Cross Correlation (NCC) algorithm (Zhao et al., 2006). Let $\{P_{1r,k},\ P_{2r,k},\ \ldots\ ,\ P_{Nr,k}\}$ denote the set of locations, output

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

from the multi-region scene matching algorithm, representing the centre coordinates of the matched sub-regions. Under an affine transformation, the location of a UAV in the 2D case can be determined by three known sub-region centre locations. In this work, more than three sub-regions in the sensed image are selected and registered with the reference image. To minimise localisation error propagated from matching errors associated with the registered locations of these sub-regions, we propose a location optimisation method which achieves a robust UAV localisation with minimum match error by comparing all possible localisation outcomes.

For practical image registration problems, in particular for the planar case, affine transformations provide a good approximation. Let the centre locations of the three sub-regions be $P_0 = (x_0, y_0)$, $P_1 = (x_1, y_1)$, and $P_2 = (x_2, y_2)$ in the sensed image, and the registered locations $P_{0r} = (x_{0r}, y_{0r})$, $P_{1r} = (x_{1r}, y_{1r})$ and $P_{2r} = (x_{2r}, y_{2r})$ in the reference image, respectively. The affine transformation from sensed image to reference image is given by

$$\begin{bmatrix} x_{0r} & x_{1r} & x_{2r} \\ y_{0r} & y_{1r} & y_{2r} \end{bmatrix} = \begin{bmatrix} a_{00} & a_{01} & b_{00} \\ a_{10} & a_{11} & b_{10} \end{bmatrix} \cdot \begin{bmatrix} x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \\ 1 & 1 & 1 \end{bmatrix} \tag{7}$$

In view of Equation (7), the affine transformation matrix is given by

$$T^{P_{0r},P_{1r},P_{2r}} = \begin{bmatrix} a_{00} & a_{01} & b_{00} \\ a_{10} & a_{11} & b_{10} \end{bmatrix} = \begin{bmatrix} x_{0r} & x_{1r} & x_{2r} \\ y_{0r} & y_{1r} & y_{2r} \end{bmatrix} \cdot \begin{bmatrix} x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \\ 1 & 1 & 1 \end{bmatrix}^{-1} . \tag{8}$$

A solution exists if these three points are not collinear (i.e., the matrix inversion exists).

Assuming the registered locations of regions P_{0r}, P_{1r} and P_{2r} are correct, i.e. the affine transformation represented by Equation (8) is correct, the locations of other sub-regions could be estimated by

$$\begin{bmatrix} x_{3e} & x_{4e} & \cdots & x_{Ne} \\ y_{3e} & y_{4e} & \cdots & y_{Ne} \end{bmatrix} = T^{P_{0r},P_{1r},P_{2r}} \cdot \begin{bmatrix} x_3 & x_4 & \cdots & x_N \\ y_3 & y_4 & \cdots & y_N \\ 1 & 1 & \cdots & 1 \end{bmatrix} \tag{9}$$

The relative estimation error of region P_3 is

$$E_{3re}^{P_{0r}P_{1r}P_{2r}} = \overline{P_{3e}P_{3r}} = \sqrt{(x_{3e} - x_{3r})^2 + (y_{3e} - y_{3r})^2} \tag{10}$$

Similarly, relative estimation errors of the other regions are expressed as follows,

$$\mathbf{E}^{P_{0r}P_{1r}P_{2r}} = \{E_{3re}^{P_{0r}P_{1r}P_{2r}}, E_{4re}^{P_{0r}P_{1r}P_{2r}}, \ldots, E_{Nre}^{P_{0r}P_{1r}P_{2r}}\}. \tag{11}$$

When the registered locations of sub-regions P_{0r}, P_{1r} and P_{3r} are assumed to be correct, the relative estimation errors are denoted by

$$\mathbf{E}^{P_{0r}P_{1r}P_{3r}} = \{E_{2re}^{P_{0r}P_{1r}P_{3r}}, E_{4re}^{P_{0r}P_{1r}P_{3r}}, \ldots, E_{Nre}^{P_{0r}P_{1r}P_{3r}}\}. \tag{12}$$

By taking all of the region-combinations under consideration, there will be C_n^3 sets of relative estimation errors; that is,

$$\{\mathbf{E}^{P_{0r}P_{1r}P_{2r}}, \mathbf{E}^{P_{0r}P_{1r}P_{3r}}, \ldots, \mathbf{E}^{P_{(N-2)r}P_{(N-1)r}P_{Nr}}\}. \tag{13}$$

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

Three regions are selected based on the criterion that the minimum relative estimation error is the lowest. Let the minimum relative estimation error be

$$\{\mathbf{E}_{\min}^{P_{0r}P_{1r}P_{2r}}, \mathbf{E}_{\min}^{P_{0r}P_{1r}P_{3r}}, \ldots, \mathbf{E}_{\min}^{P_{(N-2)r}P_{(N-1)r}P_{Nr}}\} \tag{14}$$

where, $\mathbf{E}_{\min}^{P_{ir}P_{jr}P_{kr}} = \min\{E_{s_1 re}^{P_{ir}P_{jr}P_{kr}}, \ldots E_{s_l re}^{P_{ir}P_{jr}P_{kr}}, \ldots, E_{s_{N-3} re}^{P_{ir}P_{jr}P_{kr}}\}$, $s_l = 1, \ldots, N$, $i \neq j$, $i \neq k$, $j \neq k$, $s_l \neq i$, $s_l \neq j$, $s_l \neq k$.

Then, the selected three regions are

$$\{P_{t_1 r}, P_{t_2 r}, P_{t_3 r}\} = \underset{P_{t_1 r}, P_{t_2 r}, P_{t_3 r}}{\arg\min}\ \mathbf{E}_{\min}^{P_{t_1 r}P_{t_2 r}P_{t_3 r}} \tag{15}$$

where, $t_1, t_2, t_3 \in [1, \ldots, N]$, $t_1 \neq t_2$, $t_1 \neq t_3$, $t_2 \neq t_3$.

Finally, the location P of the UAV is calculated by taking the selected three regions $\{P_{t_1 r}, P_{t_2 r}, P_{t_3 r}\}$ as the reference points as follows

$$T^{P_{t_1 r}, P_{t_2 r}, P_{t_3 r}} \cdot \begin{bmatrix} w/2 \\ l/2 \\ 1 \end{bmatrix} = \begin{bmatrix} x_{t_1 r} & x_{t_2 r} & x_{t_3 r} \\ y_{t_1 r} & y_{t_2 r} & y_{t_3 r} \end{bmatrix} \cdot \begin{bmatrix} x_{t_1} & x_{t_2} & x_{t_3} \\ y_{t_1} & y_{t_2} & y_{t_3} \\ 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} w/2 \\ l/2 \\ 1 \end{bmatrix} \tag{16}$$

where, l and w are the length and width of the sensed image respectively, i.e. $P = (w/2, l/2)$.

The scaling difference between the sensed image and the reference image is

$$b = \frac{\overline{P_{t_1}P_{t_2}}}{\overline{P_{t_1 r}P_{t_2 r}}}. \tag{17}$$

with $\overline{P_{t_1}P_{t_2}}$ and $\overline{P_{t_1 r}P_{t_2 r}}$ are the Euclidean distances.

By introducing two reference points $P'_{t_1} = P_{t_1} + (1, 0)$ and $P'_{t_1 r} = P_{t_1 r} + (1, 0)$, we compute the rotation difference between the two images:

$$\alpha = \angle P_{t_2}P_{t_1}P'_{t_1} - \angle P_{t_{2r}}P_{t_1 r}P'_{t_1 r}. \tag{18}$$

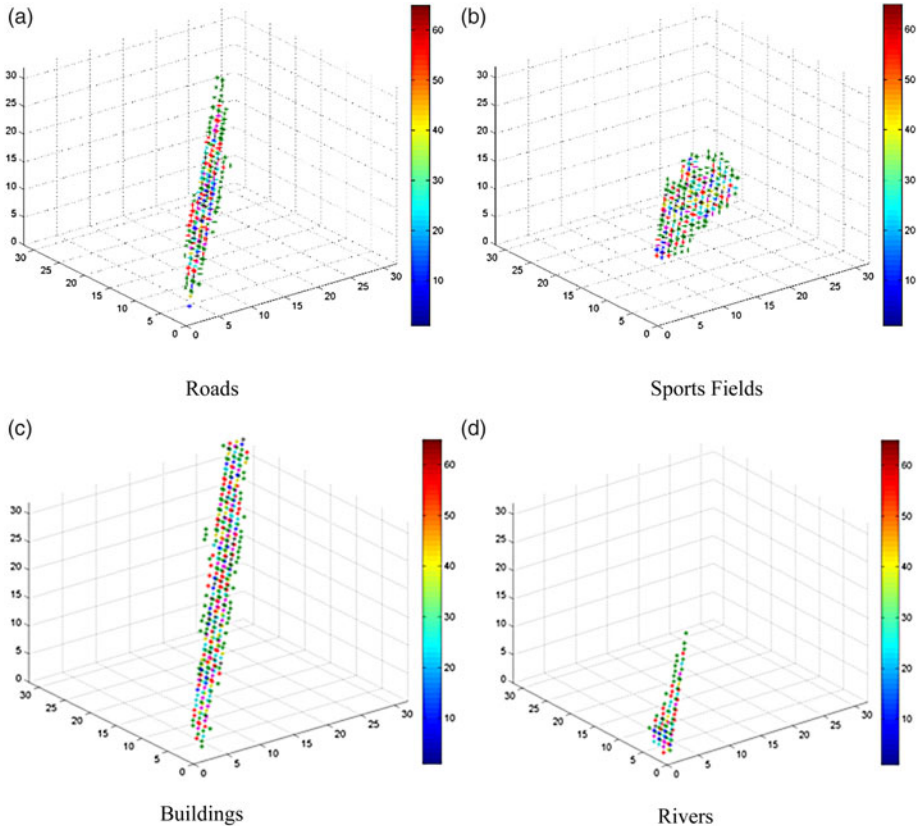
4. EXPERIMENTAL RESULTS AND OBSERVATIONS.

All of the reference images used here are taken from Google Earth. They cover both suitable scene matching areas and those regions which are difficult to match. By contaminating the reference images (i.e. adding noises, adjusting colours, rotating, zooming), 24 sensed images are obtained from the reference images for the first three experiments. These image pairs are quite different from each other, which are used to approximate real world situations for the validation of the performance of the proposed system. A number of "cutouts" of roads, sports fields, buildings and rivers from Google Earth images are used as training sets to obtain landmark histograms for each type of landmark. The colour histograms $H_{\mathbf{I}_{\text{Training}}}$ of the four landmark training sets are presented in Figure 5. These histograms are clearly distinguishable from each other.

4.1. *Colour Constancy Processing.*

Some examples of the colour constancy processing as described in Section 3.1 are illustrated in Figure 6. Column (a) is a set of "original" sensed images; column (b) shows the processing results using prior UAV location and attitude information provided by the NEC system, which shows artificial colours; images in column (c) are the results using the estimated locations and attitude information obtained after the first iteration of the proposed system, which possess

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press



Roads

Sports Fields

Buildings

Rivers

Figure 5. Colour Histograms of Training Sets.

consistent colours with the corresponding sub-regions in the geo-referenced image shown in column (d).

Figure 6 demonstrates that the colour constancy processing in the iteration circle of a UAV localisation system enhances the colour consistency between sensed image and reference image.

4.2. *Sub-region Extraction.* Based on trained colour histograms, sub-regions containing landmarks are detected as described in Section 3.2. The computed landmark likelihood maps in the first iteration for the 18th sensed image are shown in the first row of Figure 7, and the second row shows the selected landmark regions. Indicated by red rectangles, five sub-regions are selected in Figure 7(a), where the regions labelled as 2, 3, 4 and 5 contain roads and the one labelled as 1 is a false alarm. In Figure 7(b), five sub-regions which contain sport fields are detected, among which the sub-regions labelled 1 and 2 contain sports fields, but the Regions 3, 4 and 5 are false alarms. Figure 7(c) illustrates five landmark sub-regions selected, all of which contain buildings. In Figure 7(d), three regions are chosen as landmark regions of rivers. The sub-region 2 has a river, but the other two regions do not.

In the second iteration circle of the UAV localisation process, colour constancy processing is introduced. The probability maps of landmark estimation on the 18th sensed

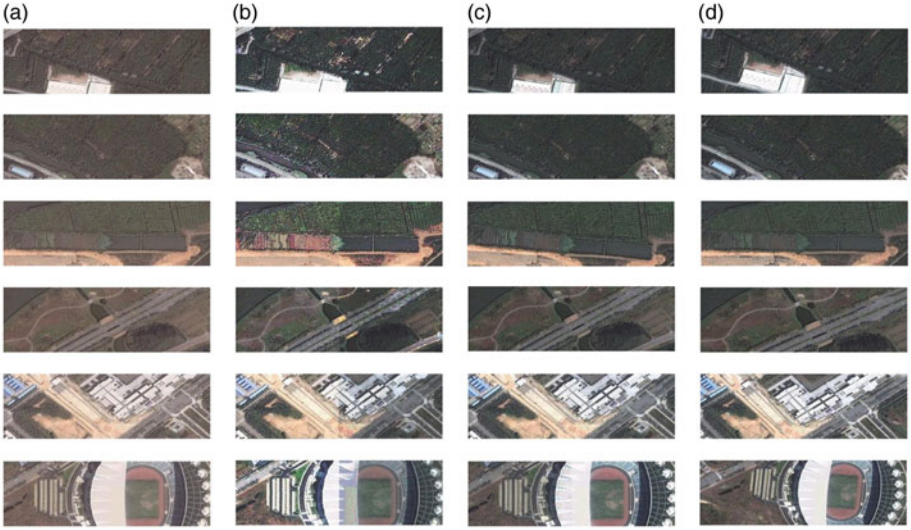


Figure 6. Colour Constancy Processing Results.

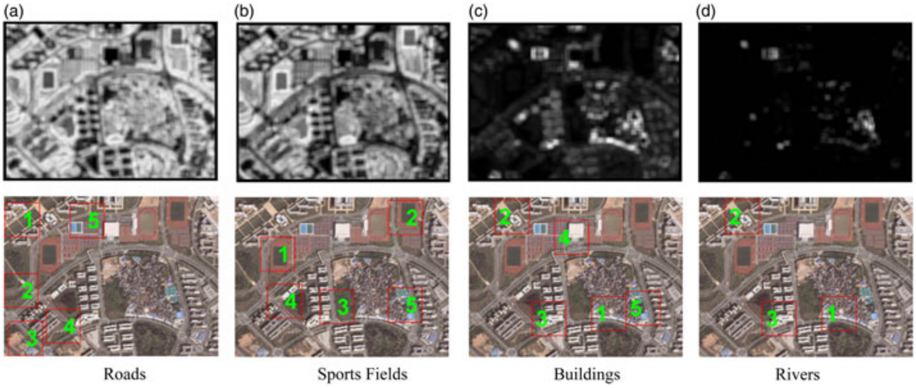


Figure 7. Landmark Estimation on 18th Sensed Image.

image are shown in Figure 8. Comparing with Figure 7, it is found that the overall landmark selection performance is improved. In Figure 8(a), all ten selected sub-regions contain the landmark of roads. In Figure 8(b), only one out of five selected regions of sports fields (i.e., Region 5) is a false alarm. A similar situation pertains in Figure 8(c), where one out of ten selected sub-regions (i.e., the ninth sub-region) does not contain building landmarks. Compared with Figure 7(d), two more sub-regions for the landmark of rivers were selected in Figure 8(d).

More examples for the landmark region selection (within the red rectangles) are given in Figure 9. Colour constancy processing in the iterative processing circle improves the performance of landmark estimation based on colour histograms and reduces false detection rates as demonstrated in Table 1.

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

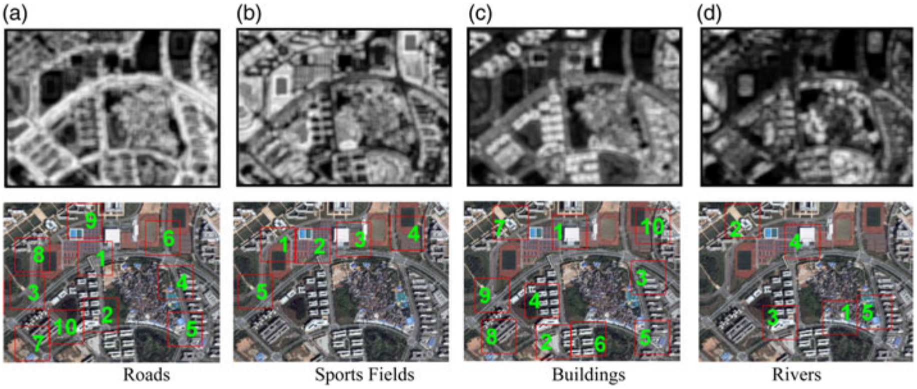


Figure 8. Landmark Estimation on 18th Sensed Images after Colour Constancy Processing.

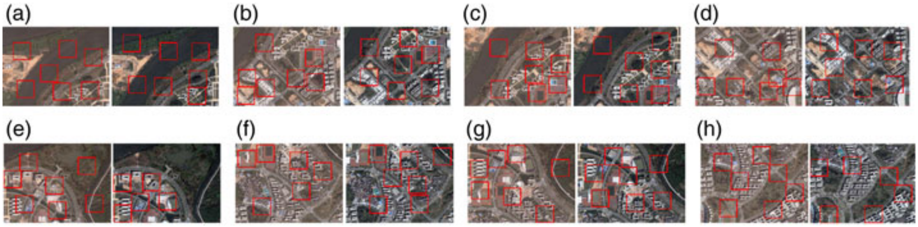


Figure 9. Eight Pairs of Landmarks detected on Sensed Images before/after Colour Constancy Processing.

Table 1. False Alarm Rates of Landmark Detection (%).

Method	Roads	Sports Fields	Buildings	Rivers
Without Colour Constancy	33·49	73·70	36·97	42·92
With Colour Constancy	15·42	67·80	17·50	39·83

4.3. *Multi-region Scene Matching Experiment.* Multi-region scene matching is performed using the NCC method (Zhao et al., 2006) in a parallel computing structure. Contaminated images from Google Earth were used and we implemented and compared the matching performance over three scene matching algorithms. We denote the proposed multi-region scene matching involving the full UAV localisation circle (see Figure 2) as MRSM-CC; and the localisation process without colour constancy processing as MRSM. A scene matching algorithm which uses the whole scene on the sensed image is also implemented and denoted by SISM (Single Image Scene Matching).

Firstly, we present the matching error comparison of localisation results using the three algorithms described above in Figure 10. It is clear that the proposed MRSM-CC algorithm outperforms the other two algorithms in terms of the localisation

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

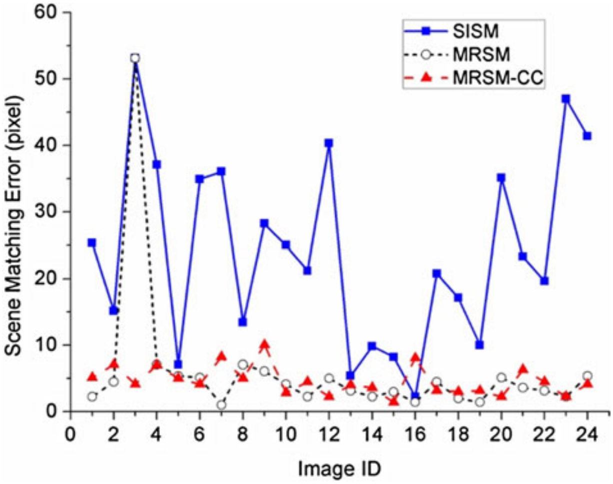


Figure 10. Comparison of matching errors.

error. All the localisation errors of the MRSM-CC are within ten pixels. While the error performance of MRSM is quite similar to that of MRSM-CC, the third scene matching result yields a large localisation error. On the other hand, the average localisation error of the SISM is significantly higher than that of the other two.

Secondly, we evaluated the proposed localisation system via Monte Carlo experiments. A total of 55 reference images were used, varying in scaling, rotation, shifting and colour. The scaling parameter between the reference image and sensed image was randomly drawn from a uniform distribution [0·9, 1·1]. Similarly, the rotation parameter was randomly drawn from [−10°, 10°]. We repeated the first experiment 55 times in the presence of randomness to the scaling and rotation parameters. Table 2 is the statistical comparison of scene matching results. An algorithm in a single run is counted as a success only if the matching error is within 25 pixels (≈10 metres).

Both Figure 10 and Table 2 show that the scene matching error performances of the proposed MRSM-CC and MRSM are significantly better than that of SISM. Clearly, the proposed MRSM-CC algorithm has a robust localisation performance under various “noise” environments. Without colour constancy processing, the MRSM may yield outliers induced by inconsistent landmark colours.

Thirdly, computational complexities of the three methods are statistically compared in terms of Central Processing Unit (CPU) time. The average size of the geo-reference images is 897 × 838 pixels and the average size of sensed images is 400 × 300 pixels. The sizes of all sub-regions are 64 × 64 pixels. All algorithms are implemented and run in Matlab. The CPU of the computer used for the experiment has eight cores, but not all cores were used. For example, if just three landmark regions are detected, the MRSM algorithm just uses three cores in scene matching. Table 3 lists the average CPU time consumed by SISM, MRSM, and MRSM-CC. Note that the CPU time of MRSM-CC is for a single iteration. When *n* iterations of the UAV localisation process are required, the CPU time is roughly *n* times. Table 3 indicates that SISM has a higher computational complexity; the CPU time consumed by MRSM is 15·07% of SISM, and by MRSM-CC is 23·73% of SISM for a single iteration.

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

Table 2. Statistical Comparison of Scene Matching (Unit: pixel).

Method	Mean Matching Error	Success Rate of Scene Matching
SISM	85·01	39·58%
MRSM	21·16	94·17%
MRSM-CC	6·31	96·67%

Table 3. CPU Time Comparison (s).

Functions	SISM	MRSM	MRSM-CC
Colour Constancy Processing	—	—	2·459
Multi-Region Extraction	—	0·481	0·481
Image Registration	28·422	3·794	3·794
Localisation via Affine Transformation	—	0·008	0·008
Total	28·422	4·283	6·742

In summary, the above experimental results indicate that the proposed localisation algorithm MRSM-CC is the most effective and efficient method over the other two algorithms examined in the performance comparison.

4.4. *UAV trajectory estimation from a sequence of sensed images.* To demonstrate the effectiveness of the proposed MRSM-CC algorithm, we carried out an experiment that estimates the trajectory of a UAV from a sequence of images taken from a Phantom 3 DJI drone as shown in Figure 11, flying over the campus of Northwestern Polytechnical University in Xi'an. Both sensed images and reference images were taken along the same path at different dates and times. The aerial imaging drone has a CMOS camera with 12 million effective pixels and can stably point straight downward during flying.

The aerial imaging drone records videos as it performs level flying along a specified path of 2·3 kilometres at a speed of 10 m/s with average height of 1200 metres. The sensed images are sampled every two seconds from a video clip of 3 minutes and 50 seconds. Image resolution is 3000 × 3000 pixels for reference images, and 1280 × 720 pixels for sensed images. The 2D rotation differences between sensed images and the reference image after pre-processing are within the ranges [−10°, 10°]. We present four image pairs in Figure 12 to show that some scene details between a sensed image and the corresponding reference image are quite different, such as the details of buildings, shadows, and colours. Clearly these differences are caused by time-varying ground objects in the same scene, the direction or colour of the sun, and the location, altitude and attitude of the camera drone.

In addition to using the NCC technique (Zhao et al., 2006) as in Section 4.3, we also considered the Speeded-Up Robust Features (SURF) method (Bay et al., 2008) for scene matching in this experiment. The performance of MRSM-CC under either NCC or SURF is compared with that of the scene matching algorithm using the whole scene on the sensed image (denoted by SISM). Statistical comparisons are shown in Table 4, where an algorithm in a single run is counted as a success if the matching error is within 30 pixels ($\leqslant$10 metres).

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

Figure 11. The Phantom 3 DJI camera drone used for data collection.

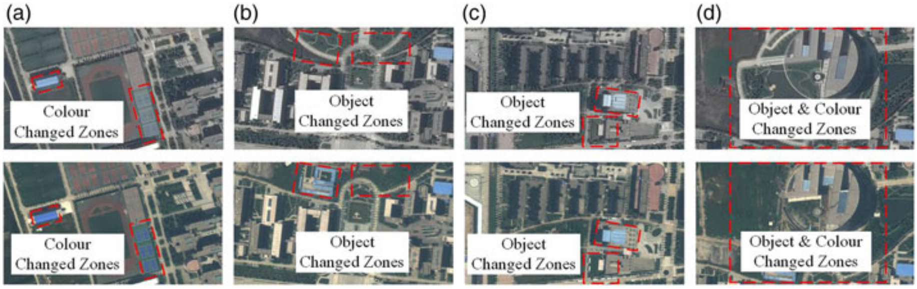


Figure 12. Four pairs of sensed images and reference images.

Table 4. Statistical Comparison of Scene Matching (Unit: pixel).

Method	Mean Matching Error	Success Rate of Scene Matching
SISM (NCC)	55·30	81·90%
MRSM-CC (NCC)	11·99	98·28%
SISM (SURF)	75·20	80·17%
MRSM-CC (SURF)	15·30	93·97%

From Table 4, it is clear that the proposed MRSM-CC method performs better than the SISM method. Because of the differences of shadows, colours, objects appearances, and shifting, scaling, 2D rotational differences between the image pair, it is difficult to match the entire scene of a sensed image with the reference image. On the other hand, using the proposed MRSM-CC system we obtained the required localisation results with an acceptable localisation error by either NCC or SURF.

In Figure 13, we present two examples that show the sensed images and registered locations in the reference image computed using different algorithms. It is observed that while the proposed statistical landmark-based matching algorithms, i.e. MASM-CC (NCC) and MASM-CC (SURF), produce acceptable results, the standard image registration approaches (via correlation or feature), i.e. SISM (NCC) and SISM (SURF), fail. We have frequently observed similar situations when some details in the

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

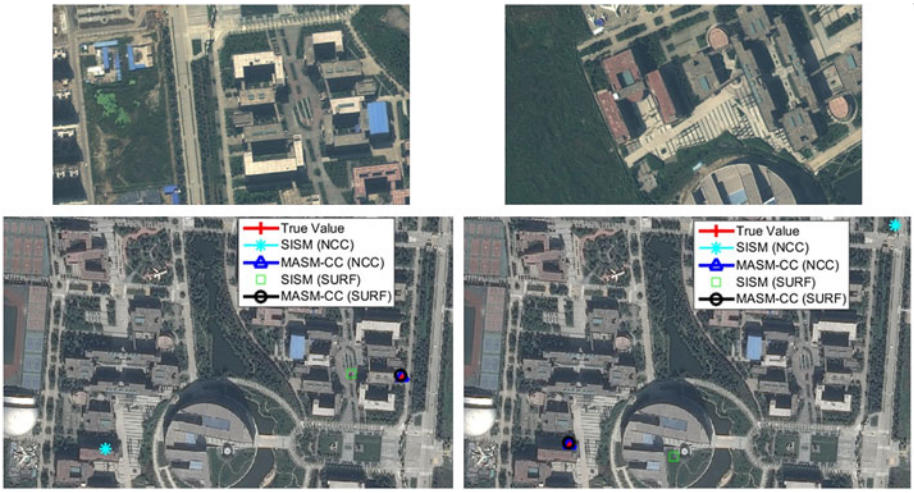


Figure 13. Two examples of sensed images and their registered locations in the reference image.

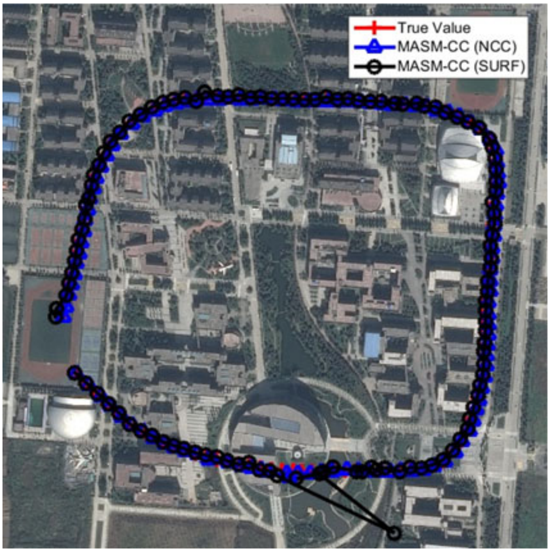


Figure 14. Trajectories of sensed images in reference image.

common scene are different between the pair of images. These two examples demonstrate the robust performance of the proposed MASM approach which uses a statistical landmark model to eliminate time-varying objects.

Figure 14 shows the estimated UAV trajectories based on a sequence of images taken by a Phantom 3 DJI drone using the proposed MRSM via NCC and SURF respectively. The trajectory estimated by MRSM-CC(NCC) is slightly better than that by MASM-CC(SURF).

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

4.5. *Observations.* Our experimental results strongly suggest that:

- The proposed MRSM-CC system is robust in two ways. First, the local EHM colour constancy processing enables sensed images to be taken in different lighting conditions from the reference image. Secondly, the use of a statistical model in landmark detection effectively eliminates time-varying objects from consideration in image registration.
- The scene matching-based UAV localisation is independently performed based on each of the sensed images. There are no accumulated errors as could be the case with INS (Wang et al., 2013), visual odometry or SLAM (Williams and Reid, 2010). In addition, localisation outliers may be eliminated via a tracking filter.
- The MRSM-CC algorithm achieves satisfactory performance using either the NCC or SURF algorithms and outperforms the SISM method.

5. CONCLUSIONS AND FUTURE WORK.

We propose a vision-based UAV localisation system where the multi-region scene matching technique under a statistical landmark region selection criterion plays a key role in registering a sensed image with a geo-referenced image. As demonstrated in our experiments, the proposed algorithm has a robust localisation performance in that: it includes colour constancy processing to eliminate the impact of colour differences between an image pair; time-varying objects can be removed from consideration by detecting landmark regions and an optimisation procedure is applied to minimise localisation error by taking into account all possible multi-region matching results using affine transformations.

It is worth mentioning that the performance of the proposed MRSM-CC algorithm also depends on the statistical model for training of landmarks. Issues associated with this vision-based localisation system, such as incorporating infrared images, and integration with other navigation sensors are under investigation by the authors.

ACKNOWLEDGEMENTS

The authors would like to thank the group of researchers in the UAV control lab at the Northwestern Polytechnical University for their assistance in the collection of the aerial data using a DJI Phantom 3 quadcopter.

FINANCIAL SUPPORT

This work was supported by the Major Program of National Natural Science Foundation of China (61135001); the National Natural Science Foundation of China (61074155); and the China Scholarship Council for one year study support at the University of Melbourne (201306290054). This work was also partially funded by Australia Research Council discovery project grant (DP120102575).

REFERENCES

Agarwal, V., Abidi, B.R., Koschan, A. and Abidi, M.A. (2006). An overview of color constancy algorithms. *Journal of Pattern Recognition Research*, **1**(1), 42–54.

Barnard, K., Cardei, V. and Funt, B. (2002). A comparison of computational color constancy algorithms -part I: Methodology and experiments with synthesized data. *Image Processing, IEEE Transactions on*, **11**(9), 972–984.

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press

Bay, H., Ess, A., Tuytelaars, T. and Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, **110**(3), 346–359.

Bonin-Font, F., Ortiz, A. and Oliver, G. (2008). Visual navigation for mobile robots: A survey. *Journal of Intelligent and Robotic Systems*, **53**(3), 263–296.

Buchsbaum, G. (1980). A spatial processor model for object colour perception. *Journal of the Franklin Institute*, **310**(1), 1–26.

Calloway, T.M., Eichel, P.H. and Jakowatz, C.V., Jr. (1990). Iterative registration of SAR imagery. *In: San Diego'90, 8–13 July, International Society for Optics and Photonics*, 412–420.

Conte, G. and Doherty, P. (2011). A visual navigation system for UAS based on geo-referenced imagery. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.* **3822**, 101–106.

Ebner, M. (2004). A parallel algorithm for color constancy. *Journal of Parallel and Distributed Computing*, **64** (1), 79–88.

Jin, Z., Pan, Q., Zhao, C. and Liu, Y. (2013). Suitability analysis based on multi-feature fusion visual saliency model in vision navigation. *In: Information Fusion (FUSION), 2013 16th International Conference on*, 235–241.

Jin, Z., Wang, X., Moran, W., Pan, Q. and Zhao, C. (2014a). Efficient scene matching using salient regions under spatial constraints. *In: Information Fusion (FUSION), 2014 17th International Conference on*, 1–8.

Jin, Z., Wang, X., Morelande, M., Moran, W., Pan, Q. and Zhao, C. (2014b). Landmark selection for scene matching with knowledge of color histogram. *In: Information Fusion (FUSION), 2014 17th International Conference on*, 1–8.

Jin, Z., Pan, Q., Zhao, C., Wei, Y. and Ma, J. (2015). Color constancy algorithm based on local exact histogram matching for scene matching navigation of UAVs. *Journal of Chinese Inertial Technology*, **23** (5), 674–680.

Jwo, D.J., Chung, F.C. and Yu, K.L. (2013). GPS/INS Integration Accuracy Enhancement Using the Interacting Multiple Model Nonlinear Filters. *Journal of Applied Research and Technology*, **11**(4), 496–509.

Li, Y., Yu, Y., Pan, Q. and Zhao, C. (2009). Scene matching based on spatial relation constraint in suitable-matching area. *In: Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*, **4**, 598–603.

Lincheng, S., Yanlong, B., Xin, X. and Liang, P. (2010). Research on matching-area suitability for scene matching aided navigation. *Acta Aeronautica et Astronautica Sinica*, **31**(3), 553–563.

Ling, Z., Pan, Q., Zhang, S., Liang, Y. and Li, Y. (2009). A scene matching method with weighted hausdorff distance based on edge measure. *Journal of Astronautics*, **4**, 053.

Lo, T.K. and Gerson, G. (1979). Guidance system position update by multiple subarea correlation. *In: 1979 Huntsville Technical Symposium, International Society for Optics and Photonics*, 30–40.

Morovic, J., Shaw, J. and Sun, P.L. (2002). A fast, non-iterative and exact histogram matching algorithm. *Pattern Recognition Letters*, **23**(1), 127–135.

Nemra, A. and Aouf, N. (2009). Robust airborne 3d visual simultaneous localisation and mapping with observability and consistency analysis. *Journal of Intelligent and Robotic Systems*, **55**(4–5), 345–376.

Siagian, C. and Itti, L. (2009). Biologically inspired mobile robot vision localisation. *Robotics, IEEE Transactions on*, **25**(4), 861–873.

Sim, D.G., Park, R.H., Kim, R.C., Lee, S.U. and Kim, I.C. (2002). Integrated position estimation using aerial image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(1), 1–18.

Van De Weijer, J., Gevers, T. and Gijsenij, A. (2007). Edge-based color constancy. *Image Processing, IEEE Transactions on*, **16**(9), 2207–2214.

Wang, T., Wang, C., Liang, J., Chen, Y. and Zhang, Y. (2013). Vision-aided inertial navigation for small unmanned aerial vehicles in GPS-denied environments. *International Journal of Advanced Robotic Systems*, **10**, 1–12.

Williams, B. and Reid, I. (2010). On combining visual slam and visual odometry. *In: Robotics and Automation (ICRA), 2010 IEEE International Conference on*, 3494–3500.

Zhao, F., Huang, Q. and Gao, W. (2006). Image matching by normalized cross-correlation. *In: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings.* 2006 *IEEE International Conference on*, **2**, II729–II732.

https://doi.org/10.1017/S0373463316000187 Published online by Cambridge University Press