Clique Descriptor of Affine Invariant Regions for Robust Wide Baseline Image Matching

Dongjoe Shin and Tardi Tjahjadi,

Abstract

Assuming that the image distortion between corresponding regions of a stereo pair of images with wide baseline can be approximated as an affine transformation if the regions are reasonably small, recent image matching algorithms have focused on affine Invariant Region (IR) detection and its description to increase the robustness in matching. However, the distinctiveness of an intensity-based region descriptor tends to deteriorate when an image includes homogeneous texture or repetitive pattern. To address this problem, we investigated the geometry of a local IR cluster (also called a clique) and propose a new clique-based image matching method. In the proposed method, the clique of an IR is estimated by Delaunay triangulation in a local affine frame and the Hausdorff distance is adopted for matching an inexact number of multiple descriptor vectors. We also introduce two adaptively weighted clique distances, where the neighbour distance in a clique is appropriately weighted according to characteristics of the local feature distribution. Experimental results show the clique-based matching method produces more tentative correspondences than variants of the SIFT-based method.

Keywords: MSER, SIFT, Affine invariant feature, Wide baseline matching, Hausdorff distance

I. Introduction

Determining correspondences between images of a scene taken at camera positions separated by a wide baseline is one of the core problems in computer vision. Although this is commonly found in many practical vision systems, it is difficult to obtain reliable results with traditional point-based matching algorithms due to the image distortion introduced by the large baseline displacement. One attempt to increase the matching reliability is to include the nearby local texture of an interesting point. For example, the Scale Invariant Feature Transform (SIFT) is designed to describe the local texture information around a scale invariant point [2] and as such it successfully improves the matching performance even when an image is degraded by noise, and by a change in scene illumination, affine transform, scale and/or 3D view point. Mikolajczyk et al. claim that SIFT-based descriptors, where a histogram of locally re-oriented image gradient is used to describe the local characteristics of a feature (e.g., SIFT, SIFT-PCA and GLOH), perform best amongst the state-of-the-art descriptors [3].

However, since the SIFT detector is originally devised for scale invariant matching, it does not fully cope with the affine distortion introduced by 3D camera motion, i.e., the matching region of an interesting point is always modelled as isotropic even if its size is adaptively determined by the scale value of the point. Consequently, the error between the estimated and actual feature points becomes larger as the baseline increases. To overcome this limitation, the shape of a local region should take into account the affine distortion involved between two views, as the projection of a planar surface is locally well modelled by an affine transform [1]. For example, the Harris affine detector [4] modifies a scale invariant region to an affine invariant region based on the fact that corresponding normalised affine regions have similarity up to 2D rotation. As a result, a shape

Dongjoe Shin and Tardi Tjahjadi are with the School of Engineering, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK.

adaptation matrix is estimated and used to transform an isotropic and scale invariant region to an anisotropic ellipse that is also invariant to scale and an affine transformation.

Other research for finding affine Invariant Regions (IR's) is motivated by wide baseline image matching where an IR is more rigorously defined as a self-adaptive image patch that automatically deforms its shape with changing viewpoint [5]. Tuytelaars et al. proposed two methods for IR detection. The first uses corners and nearby edges, and the second uses the intensity function along rays emanating from the local intensity extremum to estimate an elliptical IR [5]. A Maximally Stable Extremal Region (MSER) detector is also an intensity-based IR detector with highly desirable properties such as extremal regions are closed under continuous geometric transform and monotonic transform of image intensity [6]. Recently, the MSER detector has been extended for detecting maximally stable colour regions [7]. One advantage of using an intensity-based IR detector is that it is less complex in computation.

The most intuitive IR matching scenario is to compare the image correlations of all possible IR matches and establish Tentative Correspondences (TC's) from highly correlated IR's. Alternatively, the image correlation is often replaced by the Mahalanobis distance or multiple IR's from different scales are employed [6] to increase the possibility of having sufficient texture information. Once initial TC's have been estimated from local intensity matching, they are further refined by enforcing the global epipolar constraint within the more robust matching framework like RANSAC and MLESAC [8], [9].

In practice, to obtain more reliable initial matching results, distinctive features are extracted from an IR instead of directly using the intensity values of an IR. In addition, as distorted IRs are transformed before matching to the normalised space where two corresponding normalised IRs have similarity up to 2D rotation, a rotation invariant IR descriptor is particularly preferred. For example, Schaffalitzky et al. proposed a texture region descriptor where a rotation invariant bank of local operators represents texture regions obtained from an over-segmented image [10]. Lowe applied his SIFT descriptor to a MSER with χ^2 distance [11] because the SIFT descriptor uses local gradients which are re-oriented by locally dominant gradients. Chum et al. also proposed a non-texture based IR descriptor of a 6D-vector defined by two affine frames (called local reference and descriptor frames) and the descriptor value is used for the index value of a geometric hash table in order to perform IR matching in constant time [12].

The latest endeavour to increase the number of TC's exploits the local neighbours of an IR. The k-th nearest neighbour is used as a spatial IR proximity in a shape pair descriptor, and a pair matching distance is used to determine correspondences [11]. Thus, whenever a match is found, two pairs of IR's are added to the current set of tentative correspondences as each correspondence is supported by its closest neighbour in each view. Forssén et al. claim that the performance of this approach is normally similar to a SIFT descriptor matching but is better for images with near occlusions. In this paper, we extend the concept of the shape pair descriptor to a group descriptor, referred to as a clique descriptor in this paper. A clique descriptor is defined on an IR cluster, which consists of a seed IR and its neighbour IR's in a locally normalised Voronoi space. To ensure the robustness of a clique descriptor matching to noise, a clique Hausdorff distance which can weight neighbour distances appropriately is proposed.

This paper is organised as follows. Section 2 briefly explains the MSER detector which is used in our IR detector. The construction of a normalised IR patch and existing descriptors (e.g., SIFT and shape descriptor) are also explained. Section 3 presents the proposed clique descriptor and Section 4 presents the associated matching method. Finally, the experimental results and conclusions are presented in Section 5 and 6, respectively.

II. Invariant region detector and descriptor

A. MSER detector

The MSER detector is employed to detect IR's in the proposed matching method due to its simplicity and fast implementation (e.g., it easily detects IR's by thresholding an image). A MSER is defined by an extremal property of the intensity function of a region and its outer boundary [6]. Let $\mathrm{I}(\mathbf{p})$, where $\mathbf{p}$ is a point in an image, be a function that returns the intensity values of a set $\mathcal{I}$, e.g., an 8-bit grey level image has $\mathcal{I} = \{0, 1, 2, \cdots, 255\}$. A maximum intensity region $\mathcal{R}_m$ is then defined by

$$\mathcal{R}_m = \{\mathbf{p} \mid \mathrm{I}(\mathbf{p}) > \mathrm{I}(\mathbf{q}), \text{where } \forall \mathbf{p} \in \mathcal{R}, \forall \mathbf{q} \in \partial\mathcal{R}\}, \tag{1}$$

where $\mathcal{R}$ represents a region in an image, i.e., a set of 8-connected neighbour points, and $\partial\mathcal{R}$ is its boundary. The minimal intensity region $\mathcal{R}_n$ is defined by the opposite condition of $\mathcal{R}_m$, i.e., $\mathrm{I}(\mathbf{p}) < \mathrm{I}(\mathbf{q})$. In practice, the extremal regions in (1) are estimated by image thresholding so that a current extremal region is either split into multiple regions or merged into another region as value of the image thresholding parameter is varied, thus producing a sequence of nested extremal regions.

The MSER detector determines IR's from every sequence of nested extremal regions that satisfy the stability condition of

$$d(\mathcal{E}_i) = \frac{|\mathcal{E}_{i+\Delta}| - |\mathcal{E}_{i-\Delta}|}{|\mathcal{E}_i|}, \tag{2}$$

where $|\cdot|$ denotes the number of elements in a set and Δ is a small increment. For a sequence of nested extremal regions, $\mathcal{E}_1 \subset \mathcal{E}_2 \cdots \subset \mathcal{E}_k$, the i-th extremal region $\mathcal{E}_i$, where $(1 < i < k)$, is selected as a maximally stable extremal region when $\mathcal{E}_i$ is a local minimum of (2). Therefore, a result of MSER detection comprises binarised regions with areas that do not change significantly even when there is a small change in threshold values.

The covariance matrix of a MSER defines an elliptical IR. A maximally stable $\mathcal{E}_i$ is represented by a 2×2 matrix, $C_i = |\mathcal{E}_i|^{-1} \sum_{p\in\mathcal{E}_i} (\mathbf{p}-\mathbf{m}_i)(\mathbf{p}-\mathbf{m}_i)^T$, where $\mathbf{m}_i$ is the mean position of $\mathcal{E}_i$, T is the transpose operator, and anisotropy is measured by the ratio of two eigen values of C_i. The eigen vectors of C_i and $\mathbf{m}_i$ define a local reference frame of an IR, which is used for searching a local neighbourhood.

A MSER normalisation is a process to transform various elliptical IR's of different orientation and scale to $N_p \times N_p$ image patches for robust matching. The process is similar to a process that transforms a random data with high anisotropy ratio to one with a normalised covariance matrix $\bar{C}_i = I$. A covariance matrix can be decomposed to $C_i = U\mathrm{diag}(\lambda_1, \lambda_2)U^T$, where $UU^T = I$ and $\mathrm{diag}(\lambda_1, \cdots, \lambda_i)$ is a square diagonal matrix with elements $\lambda_1, \cdots, \lambda_i$. To make C_i isotropic, it needs to be transformed to

$$\bar{C}_i = \mathrm{diag}(\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}})U^T C_i U \ \mathrm{diag}(\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}). \tag{3}$$

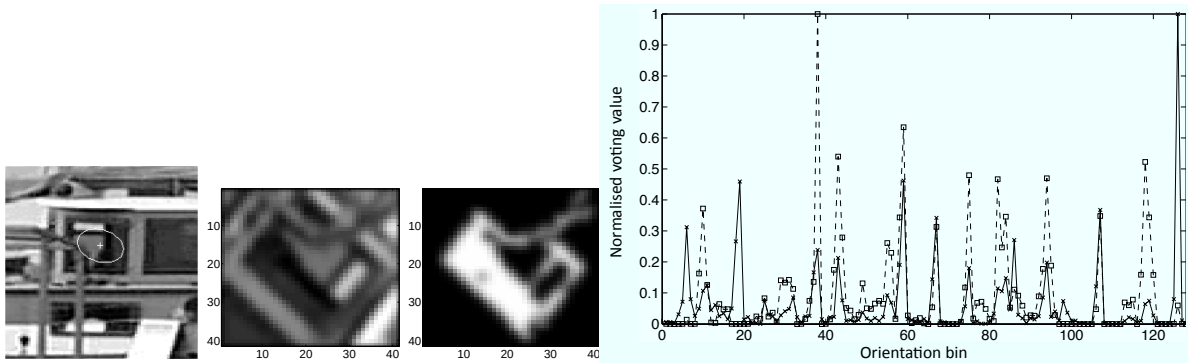
Thus, a normalised point $\bar{\mathbf{p}}$ is obtained using

$$\bar{\mathbf{p}} = s \ \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2})U^T\mathbf{p}, \tag{4}$$

where s is a scaling factor and $\mathbf{p}$ is a point belonging to an elliptical IR. A bilinear interpolation is used to estimate the intensity value of a non-integer $\bar{\mathbf{p}}$. This is followed by Gaussian blurring.

B. IR descriptor

The SIFT descriptor extracts distinctive feature vectors from the gradients of each normalised IR and we adopt the same SIFT implementation as in [2], [11]. The SIFT descriptor consists of two processes: re-orientation and local histogram estimation. In the re-orientation process, all gradient directions are re-oriented according to the dominant orientations, where the dominant



(a) (b) (c) (d)

Fig. 1. A MSER detection: (a) A detected MSER is illustrated as an ellipse and a cross denotes its centre; (b) Textured MSER of (a); (c) MSER of (a); (b) and (c) are normalised patches with $N_p = 41$; (d) SIFT description of (b) and (c) are respectively represented by a dashed line with squares and a solid line with crosses.

orientations are estimated from an orientation histogram of a normalised IR patch. The re-orientation makes the descriptor rotation invariant. Note that an IR may have multiple dominant orientations, i.e., orientations with maximum votes and with more than 80% of the votes in the orientation histogram. Thus, when estimating the orientation histogram, the orientation of a gradient is weighted by its magnitude and the value of a spatial Gaussian function centred at the normalised IR centre. In the local histogram estimation, a normalised $N_p \times N_p$ IR is divided into 16 image tiles and an 8-bin orientation histogram is estimated for each tile. Thus, a (16×8)-by-N_d histogram matrix is obtained in each normalised IR, where N_d denotes the number of dominant orientations.

The SIFT detector is normally applied to a textured IR but it may also be used with a MSER. It is thus referred to as a shape descriptor in [11]. A shape descriptor has been shown in [11] to be better than the general SIFT descriptor in matching images with near occlusions.

An example of a MSER and its normalised patches are illustrated in Fig. 1. The original image is a 800[px]×640[px] image from the Oxford data set [13] and it has been cropped to illustrate one MSER which is represented by the ellipse in Fig. 1(a). A texture region within the ellipse is normalised to give the textured MSER in Fig. 1(b) and the corresponding MSER is shown in Fig. 1(c). The scaling factor s in (4) is set to 1.2 in the normalisation so that the patches are slightly larger than the estimated IR ellipse. Also, the normalised patches are smoothed by a Gaussian kernel after a bilinear interpolation. For example, the textured MSER shown in Fig. 1 (b) has been blurred by a Gaussian function centred at the centre of the patch with a standard deviation $\sigma_t = 1$ whilst a larger standard deviation $\sigma_m = 1.2$ has been applied in Fig. 1 (c). The SIFT descriptors of the two patches are shown in Fig. 1(d), where a solid line with crosses denotes that of the textured MSER and a dashed line with squares denotes that of the shape descriptor.

III. Clique descriptor

Although the shape descriptor and SIFT descriptor perform well in the general case, the matching performance can be further improved if descriptors of nearby features of an interesting point are also used. In this context, Forssén et al. proposed a shape pair descriptor that combines the shape descriptor of the k-th nearest MSER [11]. Instead of pairwise matching, the proposed clique descriptor extends this IR grouping concept by using all the neighbours simultaneously for matching. Furthermore the neighbour distance is appropriately weighted according to local geometry and size of the elliptical MSER's.

To determine the local neighbours of a MSER $\mathcal{E}_i$, all MSER's need to be transformed to a local reference frame derived from

$\mathcal{E}_i$. This is because we assume that the configuration of locally adjacent IR's is not changed significantly in a local affine frame, and the entire feature distribution also contributes to form a local neighbour. All centres of MSER's are thus transformed to a new space defined by the centre of $\mathcal{E}_i$ and its two eigen vectors of C_i. Since this causes the selection of neighbours to be too sensitive to small variation in centre position if $\mathcal{E}_i$ has a high anisotropy ratio, MSER's with significantly small ellipses or high anisotropy ratio are excluded from the clique descriptor estimation. The transformed points are then tessellated by Delaunay triangulation, a dual of a Voronoi diagram that divides distinct n points according to the nearest neighbour rule [14].

The i-th clique is uniquely defined by a local point cluster centred at $\bar{\mathbf{m}}_i$, a normalised mean of $\mathcal{E}_i$. For example, suppose that a set of transformed MSER centres in the local reference frame of $\mathcal{E}_i$, is denoted as

$$\mathcal{V}_i = \{\bar{\mathbf{m}}_k \mid \bar{\mathbf{m}}_k = T_i\mathbf{m}_k + \mathbf{m}_i,\ k = 1, \cdots |\mathcal{E}|\}, \tag{5}$$

where T_i is the i-th affine transform which transforms the i-th MSER ellipse to a circle, $\mathbf{m}_k$ is the k-th mean of a MSER $\mathcal{E}_k$ and $|\mathcal{E}|$ is the total number of MSER's in an image. If the i-th Delaunay graph $\mathcal{G}_i$ is represented by sets of points, edges and faces, i.e., a graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{L}_i, \mathcal{F}_i)$, a point cluster called a clique is then defined as a graphical unit of $\mathcal{G}_i$. For example, the k-th clique in the i-th local frame $\mathcal{C}_i(k)$ consists of a point $\bar{\mathbf{m}}_k$ called a seed and the adjacent points directly connected to the seed are denoted by $\mathcal{L}_i$. Using the clique notation in [15], a clique centred at a point $\bar{\mathbf{m}}_k$ in the i-th local frame is

$$\mathcal{C}_i(k) = \{\bar{\mathbf{m}}_k\} \cup \{\bar{\mathbf{m}}_j \mid \forall(\bar{\mathbf{m}}_k, \bar{\mathbf{m}}_j) \in \mathcal{L}_i\}. \tag{6}$$

The proposed clique descriptor is designed to store all SIFT descriptors of MSER's in the same clique. Moreover, angles defined by every two neighbours and a seed in the local frame, and normalised size of neighbour ellipses are also stored for weighting the influence of neighbours. Thus, a clique descriptor of a MSER $\mathcal{E}_i$ has three sets: a descriptor set, an angle set and a size set.

A descriptor set $\mathcal{D}_i(k)$ of $\mathcal{C}_i(k)$ is defined by

$$\mathcal{D}_i(k) = \{F_j \mid \bar{\mathbf{m}}_j \in \mathcal{C}_i(k)\}, \tag{7}$$

where F_j is a $128 \times N_d$ SIFT descriptor matrix of a MSER $\mathcal{E}_j$ whose mean is $\mathbf{m}_j$, and its angle set is given by

$$\mathcal{A}_i(k) = \left\{\theta_j | \theta_j = \cos^{-1}\left(\frac{(\bar{\mathbf{m}}_{n1} - \bar{\mathbf{m}}_s) \cdot (\bar{\mathbf{m}}_{n2} - \bar{\mathbf{m}}_s)}{|\bar{\mathbf{m}}_{n1} - \bar{\mathbf{m}}_s||\bar{\mathbf{m}}_{n2} - \bar{\mathbf{m}}_s|}\right)\right\}, \tag{8}$$

where $\bar{\mathbf{m}}_s$ is a seed of $\mathcal{C}_i(k)$ and $\forall(\bar{\mathbf{m}}_s, \bar{\mathbf{m}}_{n1}, \bar{\mathbf{m}}_{n2}) \in \mathcal{F}_i$. Its size set is defined by

$$\mathcal{S}_i(k) = \{s_j | s_j = \frac{d_s(j)}{d_s(k)}, \forall \bar{\mathbf{m}}_j \in C_i(k) \text{ and } j \neq k\}, \tag{9}$$

where $d_s(j) = \lambda_{j1}\lambda_{j2}$ and λ_{j1} and λ_{j2} are two eigen values of C_j.

Fig. 2(a) illustrates a Delaunay graph obtained in the local reference frame of the MSER shown in Fig. 1(a). The MSER has 7 neighbours in its clique, and the textured MSER and MSER patches in the neighbourhood are shown in Fig. 2(b), where T.M. denotes a textured MSER and the seed ID is 254. The angle set describing a convex shape of a clique [as shown in Fig. 1(c)] and the normalised size of neighbour ellipses [as shown in Fig. 1(d)] are used as weighting factors.

IV. Clique distance

A distance measure is required to determine a match between two corresponding cliques with different number of neighbours, and which should be robust to the presence of any false neighbours in a clique. The Hausdorff Distance (HD) satisfies these

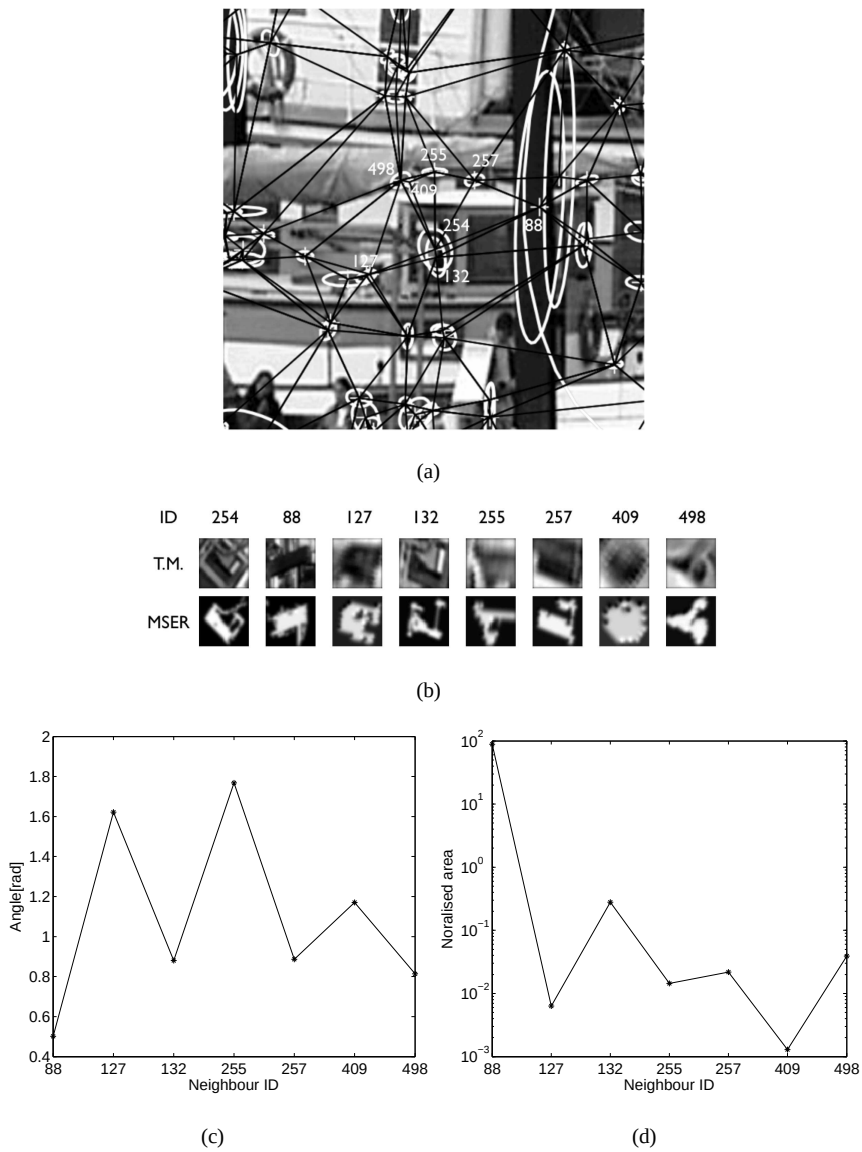


Fig. 2. A clique descriptor: (a) a Delaunay graph determined by the local reference frame of a MSER shown in Fig. 1(a); (b) 7 neighbours of seed 254 in a clique where T.M. denotes a textured MSER; (c) and (d) respectively show angle values in $\mathcal{A}_{254}(254)$ and size values $\mathcal{S}_{254}(254)$

two criteria, i.e., it defines a distance between two point sets without point correspondences and is robust against noise or outliers [16]. The general HD is a directional distance and the clique HD is

$$d_{hd}(\mathcal{C}_i(m), \mathcal{C}_j(n)) = \max_{\bar{\mathbf{m}}_\alpha \in \mathcal{C}_i} \min_{\bar{\mathbf{m}}_\beta \in \mathcal{C}_j} \{d_{\chi^2}(F_\alpha, F_\beta)\}, \tag{10}$$

where $d_{\chi^2}(\cdot)$ is a χ^2 distance that returns the minimal distance between two SIFT descriptor matrices, i.e.,

$$d_{\chi^2}(F_\alpha, F_\beta) = \min_{j,k} \frac{1}{2} \sum_{i}^{128} \frac{|F_\alpha(i,j) - F_\beta(i,k)|}{F_\alpha(i,j) + F_\beta(i,k)}. \tag{11}$$

Thus, a non-directional HD is obtained by combining two directional distances. There are various ways to combine, e.g., averaging, weighted averaging, minimum and maximum of two directional distances. However, the maximum of two directional distances is best for identification [17]. Thus, we define a non-directional HD as

$$d'_{hd}(\mathcal{C}_i(m), \mathcal{C}_j(n)) = \max(d_{hd}(\mathcal{C}_i(m), \mathcal{C}_j(n)), d_{hd}(\mathcal{C}_j(n), \mathcal{C}_i(m))).$$

If a pair descriptor is obtained by simply appending the k-th nearest neighbour to a seed descriptor and (11) is used as a metric, this distance is equivalent to the minimum between a seed distance and a neighbour distance, i.e., $\min(d_{\chi^2}(F_{s1}, F_{s2}), d_{\chi^2}(F_{n1}, F_{n2}))$. In this case the discriminant power is low because a seed distance is sometimes replaced with its closer neighbour distance. Furthermore, even though the sum of two distances is used it may be less distinctive than a single seed distance in some cases. Thus, the neighbour distance is appropriately weighted for best performance, i.e.,

$$d_w(\mathcal{C}_i(m), \mathcal{C}_j(n)) = d_{\chi^2}(F_m, F_n) + w_t d'_{hd}(N_i(m), N_j(n)), \tag{12}$$

where $N_i(m) = C_i(m) - \{\bar{\mathbf{m}}_m\}$ and w_t is a weight constant. Since (12) treats every neighbour distance equally, it is called an Equally Weighted Clique (EWC) distance in this paper and our experimental results show that EWC performs well in general.

However, since it is often unclear as to how to choose an appropriate w_t of EWC, we also propose an Adaptively Weighted Clique (AWC) distance, in which each neighbour distance has a different weighting coefficient according to the size and area of an IR, i.e.,

$$w_t = w_m \left(\frac{d_{hd}(A_i(m), A_j(n))}{a_{\max}} + \frac{d_{hd}(S_i(m), S_j(n))}{s_{\max}} \right) /2, \tag{13}$$

where w_m is a maximum neighbour weight which is normally set to 0.5, and $a_{\max}$ and $s_{\max}$ are respectively the maximum area and size distance between two images. Apart from its automatic weight selection, AWC is also advantagous particularly when the clique neighbours of a corresponding pair are changed significantly. To improve matching performance of the proposed clique matching, the general HD of (10) can be replaced with a directional clique Modified HD (MHD), i.e.,

$$d_{mhd}(\mathcal{C}_i(m), \mathcal{C}_j(n)) = \frac{1}{|\mathcal{C}_i|} \sum_{\bar{\mathbf{m}}_\alpha \in \mathcal{C}_i} \min_{\bar{\mathbf{m}}_\beta \in \mathcal{C}_j} \{d_{\chi^2}(F_\alpha, F_\beta)\}. \tag{14}$$

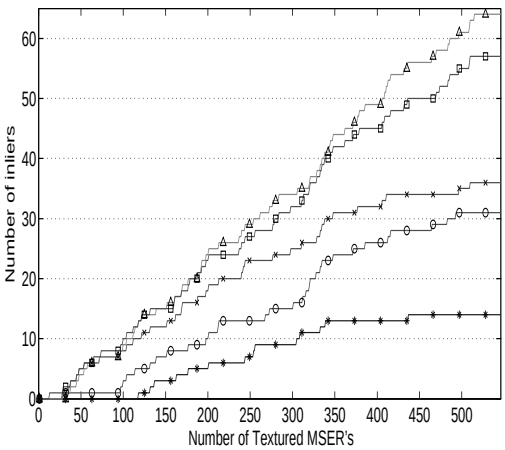
The clique matching forms TC's by collecting every matching pair for which the ratio of the best and second best clique distance is smaller than a threshold, and a neighbour IR having the smallest χ^2 distance within a clique is also included as a TC, i.e., a single matched clique produces two TC's. RANSAC is then employed to reject outliers in TC's which do not comply with the global geometric constraint such as epipolar geometry.

V. Experimental results

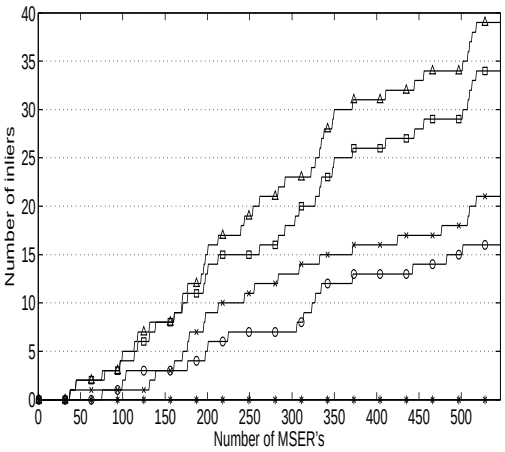
For the evaluation of matching performance, we compared three group descriptor matching algorithms, i.e., Equally Weighed Clique (EWC) matching, Adaptively Weighted Clique (AWC) descriptor matching and Pair descriptor Matching (PM). In addition, the results obtained using SIFT and Correlation Matching (CM) are presented to demonstrate the performance of IR matching with a single descriptor and without a descriptor, respectively. As a measure of matching quality, the number of inliers from initial matching results are counted and the inlier frequency graphs are estimated, where the larger the number of inliers is the better is the matching performance.

As explained in Section IV, EWC represents a clique distance where w_t of (12) is fixed, so that the contributions from all neighbour distances are equally treated when matching a seed point. However, AWC is designed to give different weight on each neighbour distance adaptively but its maximum contribution is limited by w_m which is normally set to 0.5, viz. w_t of AWC is a function of maximum weight (w_m), neighbour angle ratio (A) and neighbour size ratio (S), to be accurate. PM represents a pairwise descriptor matching method using either binarised MSERs (called a shape descriptor in [11]) or textured MSERs. The shape descriptor is considered as the SIFT result of a binarised MSER and accordingly it emphasises the shape of a MSER rather than the information from its texture, whilst a PM using textured MSERs is more concentrated on the

(a)



(b)



(c)

Fig. 3. Matching images from different views: (a) Result of equally weighted clique descriptor matching using textured MSER's; (b) and (c) Inliers frequency graphs from respectively matching using textured MSER's and binary MSER's, where: EWC descriptor (Δ), AWC descriptor ($\square$), a pair descriptor ($\circ$), SIFT (x) and correlation (*).

radiometric clues. However, both methods are similar in terms of a pairwise matching and we collectively called them PM in this paper.

In our test, a PM (e.g., either using a shape pair or texture pair descriptor) is designed to use the third nearest neighbour of a seed IR in the local reference frame to avoid a false nearest neighbour introduced by image distortion and the TC threshold is set to 1.4. This means a matching candidate of the i-th IR is selected when the ratio of the best and second best distance in the i-th row vector of a distance matrix[1] is greater than 1.4. To minimise the computational load, we do not permit an IR to have multiple correspondences so that only the candidate with the closest distance is selected as a tentative correspondence.

A. Matching images from different viewpoints

The first test compares the matching results of images from different viewpoints. When the MSER detector is applied to two 640[px]$\times$480[px] images of a tea shop as shown in Fig. 3(a), 519 and 546 MSER's are detected as input IR's. In this test, the EWC matcher with textured MSER's produces the best matching results which are illustrated as connected lines in Fig. 3(a). In a general situation (i.e., where the difference of two viewpoints is insignificant), matching based on textured MSER's will generate more TC's than matching with binary MSER's because texture in an IR gives significant clues for matching unless

[1]The size of a distance matrix is determined by $|\mathcal{E}_l| \times |\mathcal{E}_r|$ where $|\mathcal{E}_l|$ and $|\mathcal{E}_r|$ respectively represent the total number of MSER's in a left and right images. Thus, the i-th row vector includes every matching distance between the i-th MESR in a left image and any MSER in a right image.

TABLE I

INLIER RATIO OF MATCHING ALGORITHMS

	AWM	**EWM**	**PM**	**SIFT**	**CM**
${r_1}^a$	57/72[b]	64/82	31/48	36/49	14/19
${r_2}^c$	34/47	39/57	16/24	21/27	0/1
${m_r}^d$	76.5	74.1	65.3	75.0	70.0

[a]result using the textured MSER matching
[b]the number of correct matching / the number of initial TC's
[c]result using the MSER matching
[d]Average inlier ratio [%]

the texture is homogeneous or highly distorted. The two inlier frequency graphs in Fig. 3(b) and (c) illustrate the matching performance of five matching algorithms with two different input IR's. A maximum of 64 inliers are detected in the textured MSER matching whilst a maximum of 39 inliers are found in the MSER matching. Both best results as shown in Fig. 3(b) and (c) are obtained when the EWC descriptor is used. In this case, 82 and 57 TC's are estimated before RANSAC is applied, i.e., 78.0% and 68.4% of TC's are classified as inliers. Full details of the number of inliers and inlier ratios are shown in Table I.

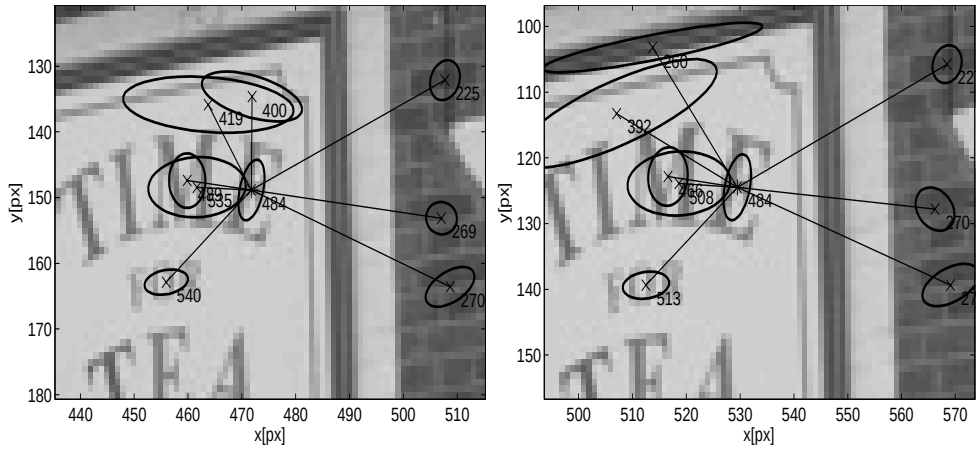
The pair descriptor performs less well than SIFT in both textured MSER and MSER matchings. The average inlier ratios of the pair and SIFT descriptor matchings are 65.3% and 75%, respectively. However, the performance is significantly degraded without a descriptor. Correlation-based matching only detects 14 inliers when textured MSER's are used (see Table I) and no inliers are found without texture information. The performance of AWC descriptor matching lies between that of the SIFT descriptor matching and EWC descriptor matching, i.e., both weighted group descriptor matchings perform better than the general SIFT descriptor matching. This result confirms that an additional neighbour distance increases the discrimination power of a single descriptor if the neighbour distance is appropriately weighted when the configuration of neighbourhood is not significantly changed.

Fig. 4 shows an example of a clique of MSER's used in matching evaluation. Although most of the neighbours support the seed MSER (ID 484), it is easily noticed that there are some false neighbours [e.g., 419 and 400 in Fig. 4(a)]. However, as the proposed method is based on the Hausdorff distance these outliers do not affect the matching result significantly. In addition, it is not essential that the number of neighbours are identical and shape of the clique may not be convex as the connectivity has been estimated from a local affine frame.

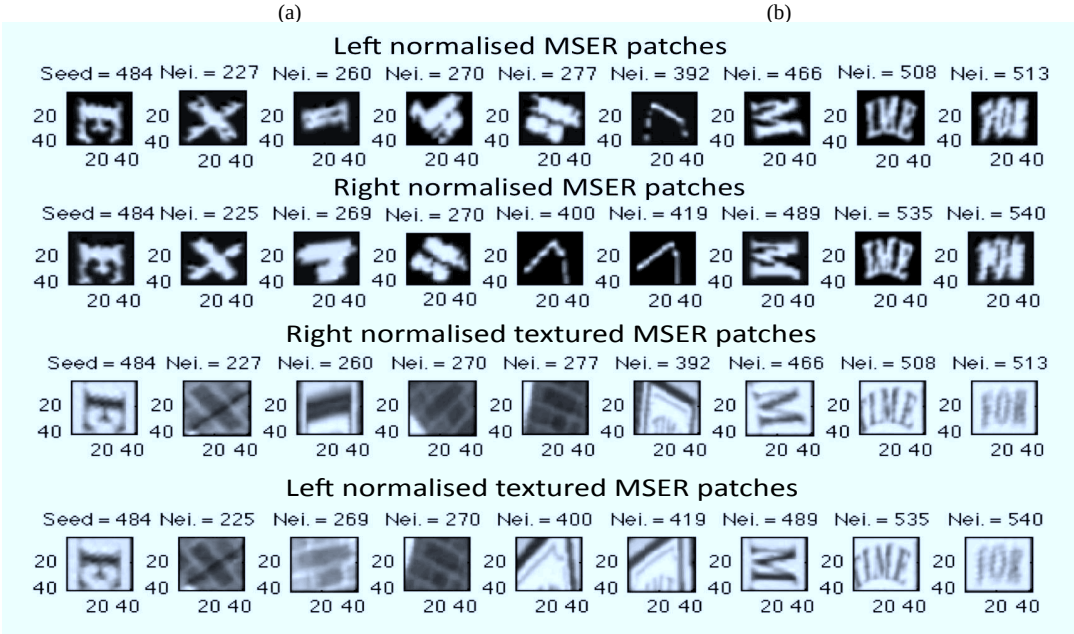
B. Matching images with repetitive pattern

The second test compares matching performances on images with a repetitive pattern, i.e., the wall images [see Fig. 5(a)] from the Oxford data set. Since the input images do not provide distinctive visual clues, matching using SIFT descriptor and correlation give similar performance, i.e., SIFT descriptor of a textured MSER is not more distinctive than a textured MSER without descriptor. However, the proposed distance improves the performance of SIFT descriptor matching as it exploits the local geometry. 1885 and 1656 MSER's are detected due to the larger size of the test images (1000[px]$\times$700[px]). However, the performance is more degraded than in the first test, e.g., the total inliers of the best matching method is reduced to less than half of the best result in the first test.

The SIFT matching of textured MSER's detects 17 inliers from 18 TC's while correlation matching detects 26 inliers out of 31 TC's, i.e., 94% and 84% inlier ratios are obtained, respectively. However, without texture information, correlation matching cannot detect any correspondence. On the other hand, since the neighbourhood does not change significantly, two weighted



(a) (b)



(c)

Fig. 4. (a) and (b) are respectively an example of left and right cliques where the seed ID is 484 and its neighbours are connected by solid lines; (c) An example of normalised MSER's used in matching: top two rows show binary MSER's and bottom two rows show textured MSER's.

neighbour distances result in the most TC's - 35 and 26 correspondences are detected out of 42 and 29 TC's by EWC and AWC, respectively [see Fig. 5(c)]. A pair distance simply adds a descriptor at the k-th nearest position from a seed IR. Thus, if the additional descriptor is not distinctive, the addition of two descriptors does not improve the matching performance. However, the EWC descriptor increases the chance of being distinctive by adding more than one neighbour description. Furthermore, the AWC descriptor penalises neighbour distance according to the shape of two matching cliques.

C. Matching images with zoom and rotation

The shape properties of a clique (such as $\mathcal{A}_i(k)$ and $\mathcal{S}_i(k)$) are not changed by camera zoom and 2D rotation because a Delaunay graph is invariant under a similarity transform, i.e., scaling, rotation and translation. Thus, the third test evaluates any effects of these camera operations on matching. An image and its zoom-out and rotated version [see Fig. 5(b)] from the Oxford data set are used. After removing small and highly anisotropic IR's, each image produces 617 and 653 MSER's. However, since the MSER detector is not scale invariant, the matching result without multi-scale MSER detection as in [11]

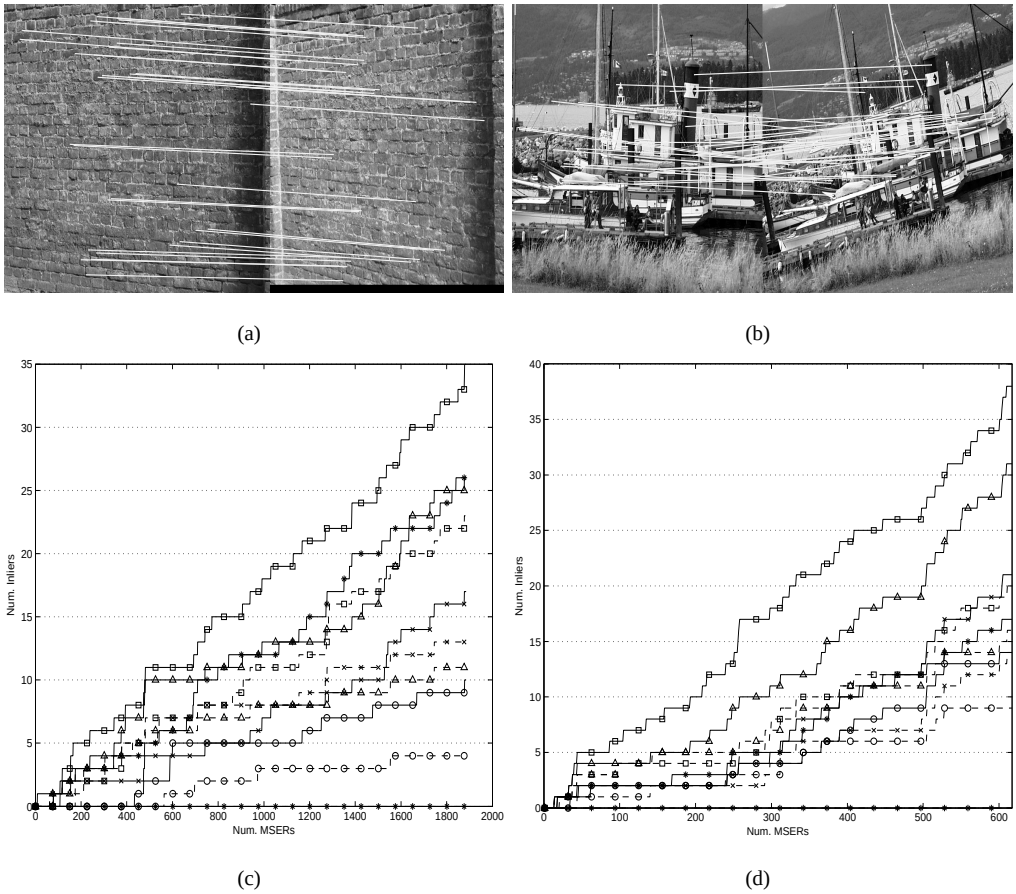


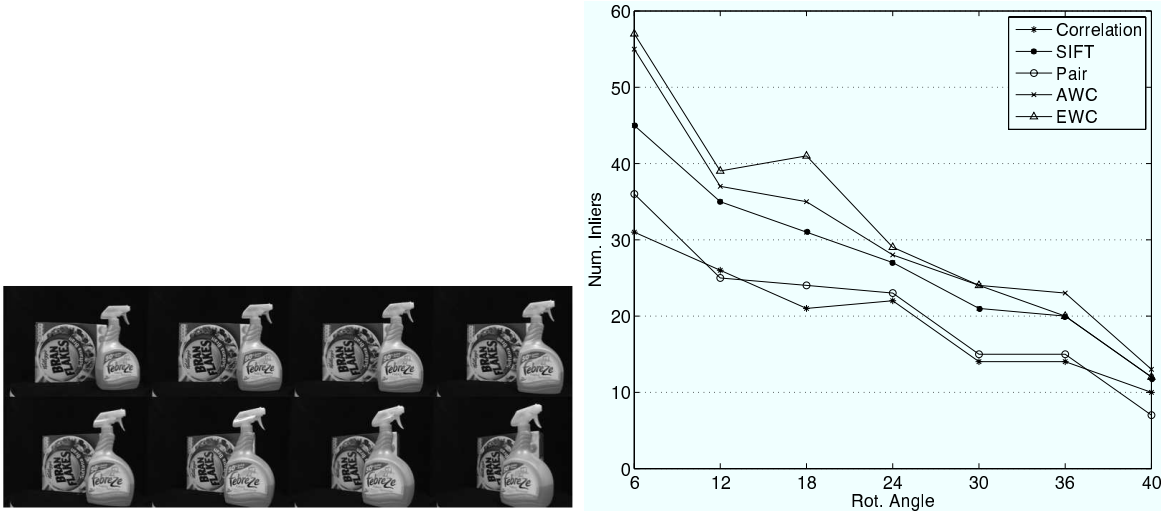
Fig. 5. (a) and (b) Results of equally weighted clique descriptor matching using textured MSER's; (c) and (d) Inlier frequency graphs of (a) and (b) respectively, where a solid and dashed lines denote matching results using textured MSER's and MSER's respectively, where: EWC descriptor (□), AWC descriptor (△), a pair descriptor (○), SIFT (x) and correlation (*).

is degraded. The EWC descriptor matching detects 38 inliers from 54 TC's, AWC descriptor matching detects 31 inliers from 43 TC's, and SIFT matching detects 21 inliers from 26 TC's [see Fig. 5(d)].

To demonstrate the performance of the proposed method under significant distortion, 5 different methods are also applied to img1.pgm (see Fig.5(b) left) and img5.pgm (not shown in the paper) of the boat image from the Oxford dataset. In this case, we used a smaller TC threshold (i.e., 1.2) to create sufficient number of TCs, and CM, SIFT, PM, EWC and AWC produced the ratios IC/TC of 5/16, 7/9, 2/8, 9/15 and 8/10, respectively. This result shows that the proposed method (i.e., EWC and AWC) performs similarly as SIFT when two images suffer from a significant distortion as the neighbour distance are set not to exceed the seed contribution.

D. Matching images with 3D camera motion

The fourth test evaluates the matching of images from a circular motion using 8 images captured at every 6° rotation from 0° to 40° as shown in Fig. 6(a). Since matching with texture information is generally better than without it, only the matching results of textured MSER's are compared. Due to the use of a black background, relatively small number of MSER's (about 130) are detected in each image. Since affine distortion is proportional to the rotation angle, the best performance of all matching methods are achieved at 6° rotation and gradually decreases as the rotation angle increases [see Fig. 6(b)]. EWC detects 56 inliers with 80% inlier ratio while SIFT matching detects 45 inliers from 56 TC's. As affine distortion increases, the performance of all methods also decreases. In particular, AWC descriptor matching detects more inliers than EWC descriptor



(a) (b)

Fig. 6. Examples of matching images generated from a circular motion: (a) images at 0° to 40°; (b) textured MSER based matching results using correlation, SIFT, Pair, AWC and EWC.

matching for rotation angle greater than 30°, which approaches the SIFT result for rotation angle greater than 36°. This is because the neighbour configuration changes significantly as rotation angle increases. However, AWC is still better than SIFT.

E. Effects of weighting coefficient

The matching distance of (12) can adjust the amount of contributions from neighbours by varying the weighting coefficient w_t from 0.0 to 1.0. For example, it gives the same distance map as SIFT when $w_t = 0$, but provides more TC's as a matched clique in EWC produces two TCs. In another extreme case when $w_t = 1$, EWC has full support from neighbours like a pair descriptor matching. However, when more than one neighbour are used in matching it is highly possible that matching candidates share neighbours from true correspondences so that it is not appropriate to set $w_t = 1$ but it is essential to ensure that the contribution of neighbours do not exceed the seed distance to avoid a false correspondence surrounded by true correspondences.

To demonstrate the effects of the weighting coefficient on matching, three more test images for each case (i.e., the cases explained in Section 5.1-5.4) are used and the number of TC's and inliers obtained using EWC are counted as w_t is varied. Figure 7(a) shows the test images, where indices A(a)-(c), B(a)-(c), C(a)-(c), and D(a)-(c) respectively represent cases of images with repetitive pattern, images from different view points, images with zoom and rotation, and images from a circular motion[2]. A solid line and a dashed line in Fig. 7(b)-(e) respectively represent the number of TC's and inliers obtained from the images in Fig. 7(a). The maximum number of TC's is normally found with small w_t (e.g., less than 0.5) and the number decreases with increasing w_t. However, these are not always the case, e.g., the maximum is with $w_t = 1.0$ for A(b) as shown in Fig. 7(b), and the number of TC's is not decreasing in D(b) as shown in Fig. 7(b).

Another interesting observation from Fig. 7(e) is that both D(a) and D(b) have the best inlier ratio around $w_t = 0.4$. This behaviour is related to the rotation angle of a circular motion as both images have been chosen with about 20 degree rotation (e.g., D(a) and D(b) in Fig. 7(a) are obtained with rotation of 20 degree and 25 degree, respectively). This is because if the rotation is too small it is difficult to demonstrate the distortion effect due to a rotation. Otherwise, overlapped regions between two images are too small for matching particularly when an image has a relatively small number of MSERs, e.g.,

[2]Test images for a circular motion are obtained from the Amsterdam Library of object images [18].

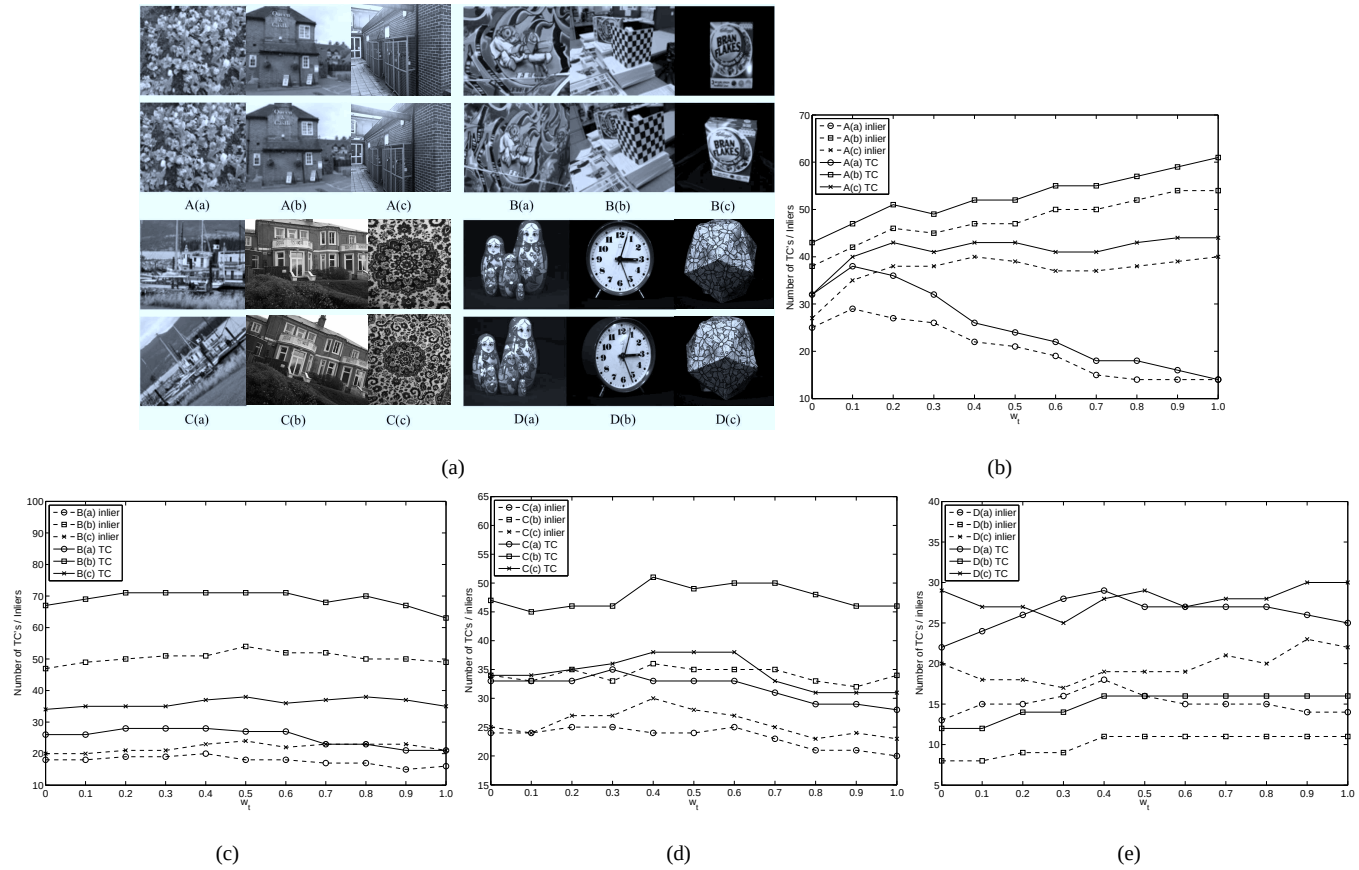


Fig. 7. Effects of weighting coefficient w_t: (a) input images are classified into four cases from A to D; (b)-(e) the number of TC's (solid line) and inliers (dashed line) obtained respectively for case A (images with repetitive pattern), case B (images from different view points), case C (zoomed and rotated images), and case D (images from a circular motion).

D(b). However, when employing an image with a small rotation (e.g., D(c) with 10 degree rotation) EWC has an inlier ratio peak at $w_t = 0.9$.

In order to compare the matching performance of five matching algorithms with textured MSERs, we present Table II with Inlier Count (IC), TC and inlier ratio of the 16 images shown in Fig. 3(a), Fig. 5(a)-(b), Fig. 6(a) and Fig. 7(a), and w_t of EWC and w_m of AWC are both fixed at 0.5. When the number of inliers is difficult to be estimated due to insufficient number of TCs (e.g., TC's < 10), we employ the fundamental matrix estimated from the best matching result and consider a TC of which the mean square error is less than 1.0 as an inlier. For example, IC of AWC of D(c) is found using a fundamental matrix estimated from EWM matching result. In terms of IC, the performance of EWC is the best, followed by AWC, SIFT, CM and PM.

To prove the hypothesis that EWC is better than the SIFT method is statistically significant, Wilcoxon signed rank test [19] was performed using the number of inliers from 16 image pairs, with the null hypothesis defined when the number of inliers from EWC and SIFT is the same. The resulting p-value is 3.052e-5 and positive and negative ranks are computed as 136 and 0, respectively. Thus, it is possible to reject the null hypothesis and say EWC can produce more inliers than SIFT at the 1% significance level.

TABLE II

INLIER COUNT (IC), TC, AND INLIER RATIO (r) OF 16 IMAGES USING 5 ALGORITHMS

	AWM[a]			**EWM**[b]			**PM**			**SIFT**			**CM**		
	IC	TC	r	IC	TC	r	IC	TC	r	IC	TC	r	IC	TC	r
A(a)	26	30	0.87	21	24	0.88	6	8	0.75	15	16	0.94	17	21	0.81
A(b)	35	38	0.92	47	52	0.90	17	21	0.81	20	26	0.77	26	28	0.93
A(c)	28	29	0.97	39	43	0.91	12	16	0.75	17	19	0.89	11	14	0.79
Fig.5(a)	26	29	0.90	35	42	0.83	10	10	1.0	17	18	0.94	26	31	0.84
B(a)	19	26	0.73	18	27	0.67	3	6	0.50	15	15	1.0	9	11	0.82
B(b)	44	60	0.73	54	71	0.76	22	43	0.51	41	52	0.79	15	23	0.65
B(c)	22	33	0.67	24	38	0.63	13	23	0.57	18	27	0.67	12	15	0.80
Fig3.(a)	57	72	0.80	64	82	0.78	31	48	0.65	36	49	0.73	14	19	0.74
C(a)	24	33	0.73	25	33	0.76	9	14	0.64	16	17	0.94	11	12	0.92
C(b)	32	43	0.74	35	49	0.71	18	27	0.67	28	33	0.85	17	24	0.71
C(d)	18	26	0.70	28	38	0.74	5	9	0.56	19	19	1.0	12	15	0.80
Fig5.(b)	31	43	0.72	38	54	0.70	14	20	0.70	21	26	0.81	17	20	0.85
D(a)	14	26	0.54	16	27	0.60	10	13	0.77	11	15	0.73	8	8	1.0
D(b)	4	8	0.50	11	16	0.69	2	6	0.33	5	9	0.56	3	6	0.50
D(c)	16	23	0.70	19	29	0.66	13	17	0.76	16	21	0.76	8	10	0.80
Fig.6(a)[c]	35	55	0.64	41	56	0.73	24	44	0.55	31	49	0.63	21	31	0.68

[a]The neighbour contribution limit is fixed at 0.5, i.e., w_m = 0.5
[b]The neighbour distance weight is fixed at 0.5, i.e., w_t = 0.5
[c]Images with 0 and 18 degree rotation are used

VI. CONCLUSION

In this paper, we explore a method that can improve the stability of wide baseline image matching. A fundamental idea behind the proposed approach relies on the assumption that the configuration of corresponding local neighbourhoods in two images from two viewpoints does not significantly change even when the two views are widely separated.

Motivated by recent research claims that affine invariant planar regions provide a strong matching clue in wide baseline image matching, the proposed method initially estimates a local affine frame from every IR and uses this information to normalise each IR in order to minimise affine distortion. After normalisation, the neighbours of an IR are estimated and used as a minimal matching unit in the proposed method. As a similarity measure of a local cluster, two matching distances (i.e., equally weighted clique descriptor distance and adaptively weighted clique descriptor distance) have been proposed and experimental results demonstrate that both distances are robust to outliers and any false neighbours in a cluster. This robustness is mainly due to the HD measure adopted in the proposed method to combine multiple descriptors of a clique set, and a detailed performance analysis of the HD measure with respect to noise level (i.e., false neighbours in a clique set) has been presented in [17]. Also, the proposed method normally outperforms the SIFT descriptor matcher and its minimum performance is bounded by the SIFT result as the neighbour contributions are adaptively weighted not to exceed the seed contribution.

However, it is worth noting that the performance of the proposed EWC is sensitive to w_t and automatic selection of the weight parameter has not been fully investigated in this paper but left as a future work. In addition, the current work has yet to investigate multi-resolution IR's, and determining neighbourhood in a locally normalised space often produces unstable results so that a more robust affine invariant frame should be explored in future research.

ACKNOWLEDGEMENT

The authors would like to thank Tardi Tjahjadi Publications for providing partial funding for the research, and all anonymous reviewers who patiently suggested many invaluable comments to enhance the quality of the paper.

REFERENCES

[1] T. Tuytelaars, L. J. V. Gool, Wide baseline stereo matching based on local, affinely invariant regions, in: Proc. Brit. Machine Vision Conf., 2000, pp. 42–56.

[2] D. G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Compt. Vis. 60 (2) (2004) 91–110.

[3] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Trans. Pattern Anal. Machine Intell. 27 (10) (2005) 1615–1630.

[4] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, Int. J. Compt. Vis. 60 (1) (2004) 63–86.

[5] T. Tuytelaars, L. J. V. Gool, Matching widely separated views based on affine invariant regions, Int. J. Compt. Vis. 59 (1) (2004) 61–85.

[6] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, Image Vision Compt. 22 (10) (2004) 761–767.

[7] P. E. Forssen, Maximally stable colour regions for recognition and matching, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[8] P. Meer, D. Mintz, D. Y. Kim, A. Rosenfeld, Robust regression methods for computer vision: A review, Int. J. Compt. Vis. 6 (1) (1991) 59–70.

[9] R. Hartley, A. Zisserman, Multiple view geometry, 1st Edition, Cambridge, 2000.

[10] F. Schaffalitzky, A. Zisserman, Viewpoint invariant texture matching and wide baseline stereo, in: Proc. IEEE Int. Conf. Computer Vision, 2001, pp. 636–643.

[11] P. E. Forssen, D. G. Lowe, Shape descriptors for maximally stable extremal regions, in: Proc. IEEE Int. Conf. Computer Vision, 2007, pp. 1–8.

[12] O. Chum, J. Matas, Geometric hashing with local affine frames, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Vol. 1, 2006, pp. 879–884.

[13] Affine covariant features, Available: http://www.robots.ox.ac.uk/~vgg/research/affine/index.html#publications [accessed on 3 Nov 2009].

[14] F. Aurenhammer, Voronoi diagrams: A survey of a fundamental geometric data structure, ACM Comput. Surv. 23 (3) (1991) 345–405.

[15] A. M. Finch, R. C. Wilson, E. R. Hancock, Matching delaunay graphs, Pattern Recognit. 30 (1) (1997) 123–140.

[16] D. P. Huttenlocher, G. A. Klanderman, W. J. Rucklidge, Comparing images using the hausdorff distance, IEEE Trans. Pattern Anal. Machine Intell. 15 (9) (1993) 850–863.

[17] M. P. D. Jolly, A. K. Jain, A modified Hausdorff distance for object matching, in: Proc. Int. Conf. Pattern Recognition, 1994, pp. A:566–568.

[18] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, The Amsterdam library of object images, Int. J. Comput. Vis., 61 (1) (2005) 103-112.

[19] D. Sheskin, Handbook of parametric and nonparametric statistical procedures, Third Edition, Chapman & Hall/CRC, 2004, pp. 189-202.