Marker Registration Technique for Handwritten Text Marker in Augmented Reality Applications

N Thanaborvorniwat\textsuperscript{1,2} and K Patanukhom\textsuperscript{1}

\textsuperscript{1}Department of Computer Engineering, Chiang Mai University, Chiang Mai, Thailand. \\
\textsuperscript{2}Graduate School, Chiang Mai University, Chiang Mai, Thailand.

*karn@eng.cmu.ac.th

Abstract. Marker registration is a fundamental process to estimate camera poses in marker-based Augmented Reality (AR) systems. We developed AR system that creates correspondence virtual objects on handwritten text markers. This paper presents a new method for registration that is robust for low-content text markers, variation of camera poses, and variation of handwritten styles. The proposed method uses Maximally Stable Extremal Regions (MSER) and polygon simplification for a feature point extraction. The experiment shows that we need to extract only five feature points per image which can provide the best registration results. An exhaustive search is used to find the best matching pattern of the feature points in two images. We also compared performance of the proposed method to some existing registration methods and found that the proposed method can provide better accuracy and time efficiency.

1. Introduction

Augmented Reality (AR) is a technology that combines vision of real-world and virtual world together by generating virtual images that are accurately overlaid with real-world scene. AR systems allow user to interact with virtual objects imposed on real environment. The AR systems have been applied in many fields such as entertainment, education, medical application, industrial design, visual art, etc. Tracking and registration of real environment’s points are significant processes to correctly overlay the virtual objects on the real scenes. The registration techniques are typically based on feature point extraction and matching procedures. Many feature points are extracted from the real scenes. The feature points in different scenes that are projected from the same 3D real-world coordinate are matched. Then camera pose in each scene can be estimated based on matching results and some additional information from depth camera [1], device orientation [2], or specific marker [3]–[5].

In this work, we develop a marker based mobile AR application that allows children to write a word in a paper to create its corresponding virtual object imposed on the paper. Fig. 1 illustrates examples of the proposed application where virtual 3D object or translated word can be created on scene images. Children can create their desired virtual objects and arrange scene composition by writing words and placing papers. The application starts from localizing texts in a scene. Then characters are recognized from those text images and are grouped into words. Each synthetic marker of the words that have virtual objects to overlay in the scene is created as a binary image of a printed word in a normal frontal view. The feature points are extracted from the markers and the texts in the scenes and are matched. Homography matrices between the marker and the scene are estimated from
the matching results. Based on estimated homography matrices, the virtual objects can be rendered into the scene image. An overview of the proposed AR application is given in Fig. 2.

![Figure 1. Example of the proposed AR application.](image1)

![Figure 2. Overview of the proposed AR application.](image2)

There are many approaches developed to detect text regions in the natural scene image [6]–[8]. In this work, we use a Maximally Stable Extremal Regions (MSER) [9] which is one of regions-based techniques implemented in many researches [7], [8]. One of advantages of the MSER based approaches is that it is invariance to affine transformation which is suitable for our application.

After the scene texts have been detection, those handwritten texts can be recognized by several methods such as using some specific features [10] or using deep-learning frameworks [11]. Some existing marker-based AR applications [3]–[5] use special pattern of markers such as four corners of a square marker [3], five shades of grayscale marker field [4] and circular marker [5] to estimate the homography. However, they cannot be used for our applications because of variation of handwritten text markers.

For registration of the printed text marker and the handwritten scene text, typical marker registration that based on general local feature matching such as SIFT [12], SURF [13], MSER [9] cannot perform well in this kind of application. The printed text marker typically consists of few characters as shown in Fig. 1 therefore it will produce less feature points and more similar descriptor for each feature point.

There are some registration approaches that focus on content specific contours such as conic shape [14] or specific characters [15]. These algorithms required specific shapes appears in target images as clues for registration which is not flexible for our application that cannot guarantee the appearance of any specific shape in the short text marker.

On the other hand, geometric rectification approaches using page layout [16] which are widely used in rectification of documents before OCR require multiple text lines or line segments in documents for analyzing vanishing points. This kind of methods is not suitable for the text markers that consist of few words or only one text line.

As a result, the registration and the homography estimation of the handwritten text markers are still a challenging problem. In this work, we develop novel approach for finding correspondences of text marker in the synthetic text marker and the real-world scene marker that is more robust to camera pose and handwritten style than the conventional methods. Most of contents presented in this paper do not focus on whole AR application but mainly focus on our proposed marker registration technique. The MSER-based text localization process applied in this work is briefly summarized in Section 2. The proposed registration method is presented in detail in Section 3. Section 4 demonstrates an experiment on the proposed method in comparison with some existing methods [7], [11], [13]. Finally, there are conclusions of this paper and direction of future works given in Section 5.

2. MSER Based Text Localization

Localization of the text in the scene image is the first process in this text-based AR application. In this work, we modified the text localization process from [8]. The MSERs can be considered as regions
with homogenous intensities that are different with their boundary intensities. The MSER can be extracted by binarizing a target image with various intensity thresholds. MSERs are some blobs at some thresholds which their areas are the most stable to the threshold variation in comparison with themselves in other thresholds. When MSERs are extracted from the image, they include text regions and non-text regions. The non-text region can be filtered out by essential shape parameters such as area, MSER intensity threshold, solidity, major axis length, minor axis length. Then, the text regions are clustered by their similarities in positions, sizes or colours. An example of MSER extraction is given in Fig. 3(a)-(c). In this work, after text regions are obtained, the result will be verified by observing surrounding areas outside the text region along its major axis to avoid missing of first or last character in the text. If there are any blobs that have equivalent properties to the characters in the text region then they will be included in that text region.

![Image](image_url)

**Figure 3.** Text localization and feature point extraction process. (a) Scene image. (b) MSERs. (c) Text mask. (d) Convex hull. (e) Examples from other camera poses. (f) Polygon simplification.

### 3. Proposed Registration Method

The proposed text marker registration for homography estimation and connected direct view of real-world are presented in the section. In the registration process, we assume that a text is already recognized, and a reference marker is already synthesized. The registration process consisted of three sub-modules that are feature point extraction, feature point matching and homography estimation as shown in Fig. 2. The proposed feature point extraction is based on contour points of a convex area of the text region and polygon simplification which provide more robustness over handwritten styles or camera poses. Centroid distances, centroid angles, and histogram of the oriented gradient (HOG) [17] are used as descriptors of the feature points. For the feature point matching process, we apply an exhaustive searching for all possible matching to find the best four matched pairs. Since there are only four to six feature points extracted in the proposed scheme, the exhaustive searching process can still perform very fast. An order consistency of the contour points is used as a constraint to avoid wrong matching results and reduce size of searching space. Details of feature point extraction and matching processes are given in the following sections.

#### 3.1. Feature Point Extraction

In order to locate the feature points, convex hulls of text regions in both reference and scene images are extracted. The convex hull is defined by a convex set of points that can visualized as the shape enclosed by a rubber band stretched around text region as shown in Fig. 3(d) where red dots are vertex point of outer contour of the convex hull. In general, we can expect that the contour’s vertex points of convex hull obtained from the same text are robust to the camera pose as demonstrated in Fig. 3(e) therefore those vertices can be used as the feature points. In order to reduce a computational cost in matching process, we reduce the number of feature points by applying a polygon simplification [18]. Unimportant points that have large angle or small segment length are repeatedly removed until the desired number of points is reached. The remaining points from reduced polygon are feature points of the text marker. An example of the final feature points is given in Fig. 3(f) and Fig. 4. Based on experiments, reducing the number of feature points not only improve the time efficiency but can also improve matching accuracy.

To estimate orientation and scale of feature points, major axes are extracted from the convex hull of the scene text and all characters that have feature points located inside. We assume that the
orientations of all feature points in the same text are approximately the same. Therefore, the orientations of all feature points are assigned as the orientation of the major axis of the convex hull of the scene text $\Phi$. However, the orientation of the major axis is unsigned, the orientation of the feature points can be $\Phi$ or $-\Phi$. The scale $\sigma$ of each feature point is assigned to be a length of the major axis of the host characters that it belongs to.

Figure 4. Feature point matching.

3.2. Feature Points Matching

In order to estimate the homography matrix that has eight unknowns, at least four corresponding points between two images are need. Let $n$ and $m$ are the number of feature points extracted from the scene image and the reference image, respectively. The parameters $n$ and $m$ are controllable in the polygon simplification step. Since the feature points are contour points which are ordering as circular sequence, an order consistency is used as a constraint for matching the feature points. As a result, there are $4 \times \frac{n!}{(n-4)4!} \times \frac{m!}{(m-4)4!}$ patterns of choosing four corresponding feature points. Fig. 4 illustrates an example of the proposed feature point matching procedure. Let $P = \{(p_i, q_i)\}_{i=1}^4$ is any possible matching pattern where $p_i$ is a feature point in the scene image and $q_i$ is a corresponding point to $p_i$ in the reference image. To find the best matching pattern, we divide the process into two stages.

The first stage is to find good candidates by using simple contour-based features that are distance and direction to a centroid of the text. Differences between the distance and the direction to a centroid of the text between two images are calculated by

$$E_d(P) = \sum_{i=1}^{4} |d(p_i) - d(q_i)| \quad \text{and} \quad E_\theta(P) = \min_{i=0} \left( \sum_{i=1}^{4} |\theta(p_i) - \theta(q_i)| \pm \Phi \right),$$

where

$$E(P) = \frac{E_d(P)}{\max_{P} \left( E_d(P) \right)} + \frac{E_\theta(P)}{\max_{P} \left( E_\theta(P) \right)}.$$

All possible matching patterns are ranked by $E(P)$. The top $K$ matching patterns that provide less $E(P)$ are selected as candidates. We also consider the distribution of the matched feature points. In order to reduce an error in the homography estimation, the matched feature points should be distributed in all four quadrants divided by the major and the minor axes in both scene marker and

---

**Figure 4. Feature point matching.**
reference marker. The candidates that has high and their feature points does not distribute in all quadrants are removed. This filtering process significantly reduces the computational cost.

The second stage implements the HOG [17] to extract descriptors of feature points. HOG descriptors of windows surrounding every feature points and HOG descriptors of every host character that have feature points are extracted. An example of all windows and host characters of the feature points is illustrated in Fig. 4. Sizes of the windows are relative to scale $\sigma$ of each feature point. In this paper, we use the windows size of $0.25\sigma$ in the experiment. While orientations of the windows and the host characters are selected from $\pm \Phi$ that provide the minimum value in the calculation of $E_\theta(P)$. Final descriptor of each feature point is obtained by concatenating the HOG descriptor from its local windows and the HOG descriptors of its host character. The best matching pattern is defined as

$$P_{\text{opt}} = \arg \min_{P} \left( \sum_{i=1}^{4} |h(p_i) - h(q_i)| \right)$$

(3)

Where $h(\cdot)$ represents the final descriptor of the corresponding feature point.

4. Experimental results

In this section, we conduct an experiment to evaluate performance of the proposed registration scheme. The experiment is conducted on 1,560 scene images of multiple views of 120 handwritten markers from 12 English words with 10 writing styles. Examples of test images are given in Fig. 5. The resolution of the scene images is 750x1000 pixels. 120 reference text images are generated by using Calibri font in normal font style. The homography matrices are estimated for every test images. Accuracy of homography estimation and registration time and homography estimation time consumption (excluding text localization) are measured. Accuracy of homography estimation is defined as a percentage of the scene markers that are correctly rectified into their frontal view. We compare performance of the proposed methods with three conventional approaches that are MSER [7], SURF [13], and contour correspondences [15]. In the baseline MSER approach, the MSERs are extracted and HOG descriptor are used for matching the MSERs in both images. Centroid positions of MSERs use as positions of the feature points for find homography. Random Sample Consensus (RANSAC) [19] is applied for finding the best homography matrix in all three conventional approaches.

![Figure 5. Examples of testing images.](image)

| Table 1. Performance comparison | Table 2. Accuracy of the proposed method when the number of feature points is varying |
|---------------------------------|---------------------------------|
| Methods                        | Accuracy [%] | Time consumption a |
| Proposed ($n=m=5$)             | 97.37        | 1.0000             |
| MSER + HOG [7]                 | 60.96        | 1.4389             |
| SURF [13]                      | 34.81        | 2.4584             |
| Contour                        | 49.74        | 1.7972             |
| Contour Correspondence [15]    |              |                    |

The performances of four methods are demonstrated in Table 1. In order to ignore the hardware specification, time consumptions are presented as ratios of average time usage in each method.
comparing with the proposed method. In the proposed method, we choose the number of feature points as five because it provides the best accuracy comparing with the other number of feature points as shown in Table 2. Examples of the results from four methods are given in Table 3 that include visualization of feature point matching results, rectified scene markers by the estimated homography (expected to be the frontal view), and the reference markers rendered to the real scene. According to the results, the proposed methods can provide the highest accuracy and the least time consumption. The proposed method extracts only few feature points which mostly appear in both images. The proposed matching scheme uses the order of the circular sequence as matching constraint to avoid miss-matching and keeps the matched feature points over four quadrants of the marker to reduce the error of homography estimation. The simple contour-based features used for rejecting bad choices also helps to improve time efficiency. The SURF method obtains the lowest accuracy and time efficiency in this dataset because SURF method extracts many feature points which have no correspondence in the other image. In addition, the binary text markers that consists of only simple lines, curves and corners produce descriptors that are non-unique in many feature points resulting in incorrect matching as shown in Table 3.

On the other hand, the conventional MSER and the contour correspondence methods can provide more robust feature points and faster than SURF. For conventional MSER method, the number of feature points is typically equal to the number of character in the text marker therefore it cannot be used for the texts that consist of less than four characters. A problem also occurs when the text markers consist of the same character more than one. The homography estimation using the MSER’s centroids produces more error than the proposed method which can be observed from the rectified scene texts in Table 3 because the character centroids are almost aligning in a straight line that may result in nearly ill-condition in the homography estimation.

For the contour correspondence method, the feature points distribute better than the conventional MSER method. However, there are still many wrong correspondences when using feature based on cross-ratio of areas.

5. Conclusion

In this paper, we presented the marker registration technique for AR application that is designed for dealing with the simple and low-texture handwritten text markers. The MSER and the contour simplification are used for extracting the feature points. The new feature point matching procedure is developed. Based on the experiment from 1,560 scene images with handwritten text markers, the proposed method provides higher accuracy, more robustness to handwritten styles and camera poses and requires less computational cost than the baseline methods.

Table 3. Examples of the results from four methods.
References

[1] Kim B S, Xu S, and Savarese S 2013 Accurate localization of 3D objects from RGB-D data using segmentation hypotheses Proc. IEEE Comput. Soc. Conf. Conf. Pattern Recognit. pp 3182–89.

[2] Kurz D and Benhimane S 2011 Gravity-aware handheld augmented reality 2011 10th IEEE nt. Symp. Mix. Augment. Reality, ISMAR 2011 pp 111–20.

[3] Park J, Jiang B, and Neumann U 1999 Vision-based pose computation: robust and accurate augmented reality tracking Proc. 2nd IEEE ACM Int. Work. Augment. Reality pp 3–12.

[4] Herout A, Szentandrasi I, Zacharia M, Dubsk a M, and Kajan R 2013 Five shades of grey for fast and reliable camera pose estimation Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp 1384–90.

[5] Pagani A, Koehle J, and Stricker D 2011 Circular Markers for camera pose estimation 12th Int. Work. Image Anal. Multimed. Interact. Serv. - WIAMIS 2011.

[6] Zhang H, Zhao K, Song Y Z, and Guo J 2013 Text extraction from natural scene image: A survey Neurocomputing 122 pp 310–23.

[7] Turki H, Halima M B, and Alimi A M 2016 Scene text detection images with pyramid image and MSER enhanced Int. Conf. Intell. Syst. Des. Appl. ISDA pp 301–6.

[8] Intaratat S and Patanukhom K 2017 Self-learning Structure for Text Localization 2017 15thIAPR International on Machine Vision Applications 2 pp 370–73.

[9] Matas J, Chum O, Urban M, and Pajdla T 2002 Robust Wide Baseline Stereo from Maximally Stable Extremal Regions Br. Mach. Vis. Conf. pp 384–93.

[10] Mingqiang Y, Kidiko Y, and Joseph R 2008 A Survey of Shape Feature Extraction Techniques,” Pattern Recognit. Tech. Technol. Appl.

[11] Cireșan D C, Meier U, Gambardella L M, and Schmidhuber J 2011 Convolutional neural network committees for handwritten character classification Proc. Int. Conf. Doc. Anal. Recognition, ICDAR 10 pp 1135–39.

[12] Lowe D G 2004 Distinctive image features from scale-invariant keypoints Int. J. Comput. Vis.60 pp 91–110.

[13] Xu A and Namit G 2008 SURF : Speeded - Up Robust Features Eur. Conf. Comput. Vis. pp 1–30.

[14] Kumar P, Jawahar C V, and Narayanan P J 2004 Geometric Structure Computation from Conics Proceeding Indian Conf. Comput. Vis. Grap hiccs Image Process. pp 1–6.

[15] Kumar M P, Jawahar C V, and Narayanan P J 2004 Building blocks for autonomous navigation using contour correspondences Image Process. 2004. ICIP ’04. 2004 Int. Conf. 2 pp 1381–84.

[16] Clark P and Mirmehdi M 2001 Estimating the Orientation and Recovery of Text Planes in a Single Image Proc. Br. Mach. Vis. Conf. 2001 pp 44.1.,-10.

[17] Dalal N and Triggs B 2005 Histograms of oriented gradients for human detection Proc. 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, CVPR 2005 I pp 886–93.

[18] Veltkamp R C 2003 On the Implementation of Polygonal Approximation Algorithms Ovidiu Grigore.

[19] Fischler M and Bolles R C 1981 Random Sample Consensus: A Paradigm for Model Fitting with Commun. ACM 24 pp 381–95.