CCD: Connected Component Descriptor for Robust Mosaicing of Camera-Captured Document Images

T Kasar and A G Ramakrishnan
Medical Intelligence and Language Engineering Laboratory
Department of Electrical Engineering, Indian Institute of Science
Bangalore, INDIA
{tkasar, ramkiag}@ee.iisc.ernet.in

Abstract

We propose a robust method for mosaicing of document images using features derived from connected components. Each connected component is described using the Angular Radial Transform (ART). To ensure geometric consistency during feature matching, the ART coefficients of a connected component are augmented with those of its two nearest neighbors. The proposed method addresses two critical issues often encountered in correspondence matching: (i) The stability of features and (ii) Robustness against false matches due to the multiple instances of characters in a document image. The use of connected components guarantees a stable localization across images. The augmented features ensure a successful correspondence matching even in the presence of multiple similar regions within the page. We illustrate the effectiveness of the proposed method on camera captured document images exhibiting large variations in viewpoint, illumination and scale.

1 Introduction

Document image analysis often requires mosaicing when it is not possible to capture a large document at a reasonable resolution in a single exposure. Such a document is captured in parts and mosaicing stitches them into a single image. There are two main approaches to image mosaicing, namely direct method and feature-based method. Though direct methods yield a dense correspondence and are very accurate, feature-based techniques are preferred since they are more robust to large geometric and photometric distortions and are also potentially faster. The success of feature-based methods depend on the stability and discriminative power of the features used. For scanned document images, it is relatively easier to establish feature correspondence because the images are uniformly illuminated and differ only by a 2-D Euclidean transformation. Camera-captured images are characterized by non-uniform illumination, blur and perspective distortion, which pose a great challenge to reliable feature extraction as well as matching. The general approach to feature matching is to first compute a set of putative matches and then use multiple view geometric relations or geometric filtering based on the local spatial arrangement of the features to disambiguate matches. It works well as long as the putative matches have a good percentage of correct matches. This is not so for document images in general. Feature matching in document images often leads to gross errors due to the multiple occurrences of the same letters or words.

2 Background

Corners have been widely used as features owing to their 2-D structure that provides the maximum information content. Harris and Stephens [4] developed a corner detector which is robust to changes in viewpoint and illumination but is sensitive to scale. Lowe [10] proposed a scale invariant feature transform (SIFT) descriptor computed from the spatial distribution of image gradients. Mikolajczyk et al [12] compared the state of the art affine covariant region detectors viz Harris affine, Hessian affine, maximally stable extremal regions, intensity-based region detector, edge-based region detector and salient regions. The outputs of these affine region detectors are described using SIFT and their performances evaluated against viewpoint changes, scale changes, blur and JPEG compression artifacts. The results largely depend on the type of scene used for the experiments, with none of the detectors clearly outperforming the others for all types of scenes and transformations. Viewpoint change was found to be the most difficult type of transformation to cope with, followed by scale change.

In [11], a host of local feature descriptors are evaluated and the SIFT descriptor is reported to give the best perfor-

mance. The SIFT descriptor has been successfully used by many researchers in a number of applications such as object recognition, generating panoramic images and image retrieval. However, being a local feature descriptor, the SIFT descriptor does not work well when there are repetitive structures in the image [9, 13]. It does not make any distinction between different instances of the same letter in the document image; this can lead to a large number of outliers in correspondence matching.

Several methods have been designed specially for document images. Wichello and Yan [16] proposed a simple method for mosaicing binary documents using a cross-correlation match. It was assumed that apriori knowledge of image placement and overlap are available. It was also assumed that there is no warping thus limiting its use to scanned documents only. Pilu and Isgrò [14] introduced a two-stage approach for mosaicing scanned documents using a corner detector described in [15] called SUSAN (Smallest Univalue Segment Assimilating Nucleus). They used an intensity-based cross-correlation technique to compute an initial transformation hypothesis, which is further used to gather more supporting matches.

Zappala et al [17] proposed a mosaicing technique where the user slides the paper to be mosaiced under a stationary, over-the-desk camera until the whole document have passed through the field of view of the camera. In their method, first the skew is corrected and then the image is segmented into a hierarchy of columns, lines and words. Point correspondences are then established by matching the lower right hand corners in pairs of overlapping images. Lian et al [9] have proposed a 2-step approach for mosaicing without restricting the motion of the camera, thus allowing greater flexibility than scanner-based or fixed-camera-based approaches. Firstly, perspective distortion and relative rotation are removed by mapping the vanishing points of text line direction and vertical character stroke directions to points at infinity. Then, PCA-SIFT is employed to establish feature correspondence. Finally, accurate registration is obtained by a cross-correlation block matching. However, segmentation of lines, columns and words is not a trivial task for complex documents.

Unlike the above methods, we use features derived from connected components (CCs) that can be easily computed in any document image irrespective of the page layout. The use of CCs is a natural choice for localization of 'interest points/regions' in document images since they are highly stable, unaffected by rotation, scale and other deformations. We use an augmented feature matching scheme for resolving ambiguities that can occur locally due to multiple similar regions in the document image. The new method is discussed in detail in the following section. Feature matching and the subsequent results of mosaicing are presented next. The conclusions are given in the end.

3 CCD: Connected Component Descriptor

We introduce a new region descriptor derived from connected components. The key advantage of the new descriptor is that the same CC can be detected across different images of a document captured under different viewing conditions. Thus, the new descriptor inherently has an excellent repeatability rate, which is highly desirable for image matching.

3.1 Localization of Measurement Regions

We use a robust method [7] of obtaining CCs from the edge image that can handle characters having different polarity of the foreground-background intensities. Canny edge detection [2] is performed individually on each channel of the color image and the edge map is obtained by combining the three edge images. A connected component labeling follows the edge detection step and the associated parameters such as convex hull and area are computed. We make some sensible assumptions about the document to remove unwanted components. The aspect ratio is constrained to lie between 0.1 and 10 to eliminate highly elongated regions. Components whose area is less than 6 pixels and larger than $1/6^{\text{th}}$ of the image dimension are eliminated. Since edge detection yields both the inner and outer boundaries of the characters, it suffices to describe only the region bounded by the outer boundary. We filter out all the inner boundaries while retaining the outer boundaries of the characters.

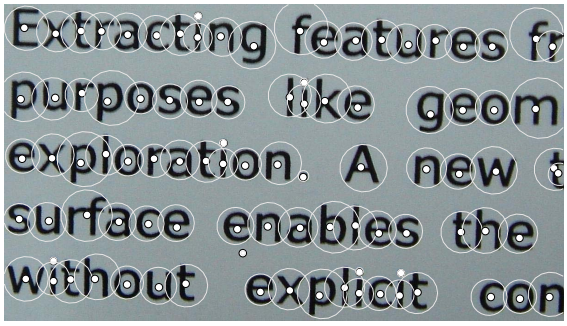


Figure 1. Measurement regions computed from each connected component.

We compute the centroid of the convex hull of each CC. The measurement region is identified as the smallest circle, centered at the computed centroid, that just encloses the CC as shown in Fig. 1. All the measurement regions are then normalized to a standard size for feature extraction.

3.2 Extraction of Invariant Features

Though the feature descriptor could be any invariant descriptor, we have chosen angular radial transform [1, 8] because of its desirable properties like compact size, robustness to noise, scaling and deformation, invariance to rotation and ability to describe complex objects.

3.2.1 Angular Radial Transform

The angular radial transform is a complex orthogonal transform defined on a unit disk and is used in MPEG-7 for shape coding. The basis functions of ART are orthogonal and hence have no redundancy. The ART coefficients of order m and n are defined by

$$\mathbf{F}_{m,n} = \int_0^{2\pi} \int_0^1 \mathbf{B}_{m,n}(\rho,\theta)\, \mathbf{I}(\rho,\theta)\, \rho\, d\rho\, d\theta \qquad (1)$$

where $\mathbf{I}(\rho,\theta)$ is the image function in polar coordinates, $\mathbf{B}_{m,n}(\rho,\theta)$ is the basis function of order m and n of ART. These ART basis functions are defined in polar coordinates and are separable along the radial and angular directions.

$$\mathbf{B}_{m,n}(\rho,\theta) = \mathbf{R}_m(\rho)\, \mathbf{A}_n(\theta) \qquad (2)$$

where m and n are non-negative integers. The radial polynomial is defined by a cosine function as follows:

$$\mathbf{R}_m(\rho) = \begin{cases} 1 & m = 0 \\ 2\cos(\pi m \rho) & otherwise \end{cases} \qquad (3)$$

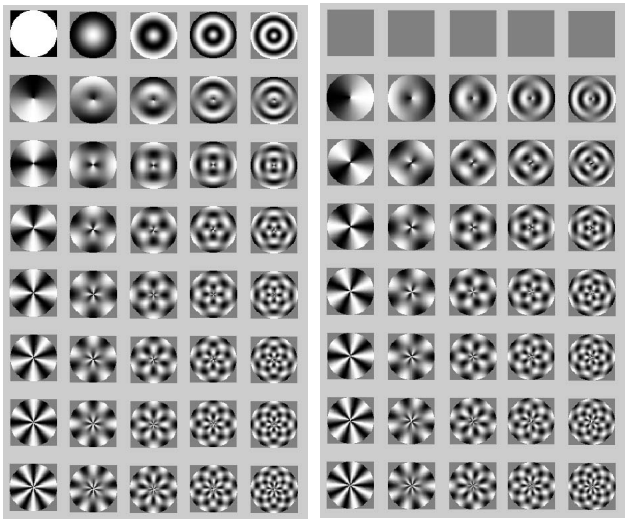
For the angular basis function, an exponential function is used to achieve invariance to rotation.

$$\mathbf{A}_n(\theta) = \frac{1}{2\pi} \exp(jn\theta) \qquad (4)$$

Since the ART basis functions exhibit symmetry and antisymmetry with respect to the x and y axes and the origin, it is shown in [6] that the complete basis functions can be obtained using only the first quadrant of the basis functions. Thus, the ART coefficients can be efficiently computed using Eqn. 5 when n is even and Eqn. 6 when n is odd.

$$\begin{aligned}\mathbf{F}_{m,2k} = \sum_{(x,y)\in\mathcal{R})} \{&[\mathbf{I}(x,y)+\mathbf{I}(-x,y)+\mathbf{I}(-x,-y)+\mathbf{I}(x,-y)]\mathbf{B}^r_{m,2k}(x,y) \\ &-j[\mathbf{I}(x,y)-\mathbf{I}(-x,y)+\mathbf{I}(-x,-y)-\mathbf{I}(x,-y)]\mathbf{B}^i_{m,2k}(x,y)\}\end{aligned} \qquad (5)$$

$$\begin{aligned}\mathbf{F}_{m,2k+1} = \sum_{(x,y)\in\mathcal{R})} \{&[\mathbf{I}(x,y)-\mathbf{I}(-x,y)-\mathbf{I}(-x,-y)+\mathbf{I}(x,-y)]\mathbf{B}^r_{m,2k+1}(x,y) \\ &-j[\mathbf{I}(x,y)+\mathbf{I}(-x,y)-\mathbf{I}(-x,-y)-\mathbf{I}(x,-y)]\mathbf{B}^i_{m,2k+1}(x,y)\}\end{aligned} \qquad (6)$$



(a) (b)

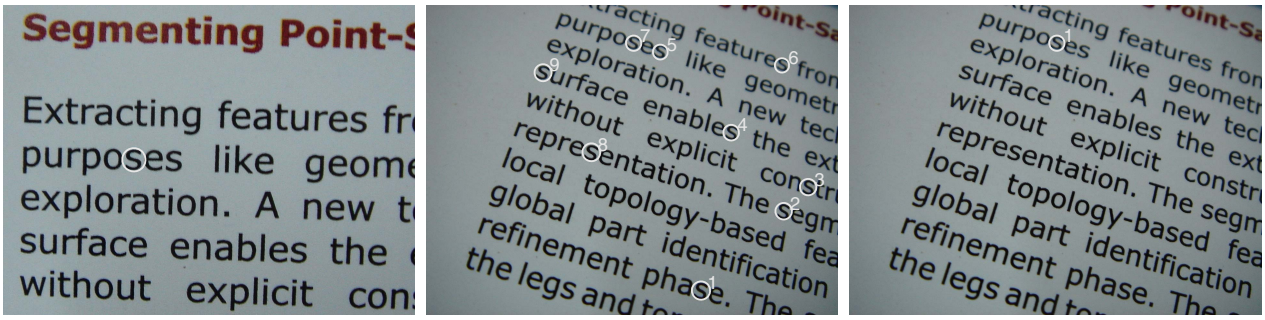
Figure 2. The (a) real and (b) imaginary part of the ART basis functions of order m = 0,1,. . . , 4 and n = 0,1, . . . , 7.

where $\mathbf{I}(x,y)$ denotes the intensity of the measurement region defined by each CC and the region of summation $\mathcal{R} = (x^2+y^2) \le 1, x \ge 0, y \ge 0$. $\mathbf{B}^r_{m,n}$ and $\mathbf{B}^i_{m,n}$ denote the real and imaginary parts of the ART basis functions of order m and n respectively. We compute the ART coefficients of order m = 0, 1, ... , 4 and n =0, 1, ... , 7. To achieve rotational invariance, the magnitude of the ART coefficients are used as the feature vector. The coefficient $\mathbf{F}_{0,0}$ is used as a scale normalization factor yielding a 39-dimensional vector for each CC.

3.3 Augmented Feature for Matching

For every CC identified in the reference image, we seek the best match in the target image and vice versa. Because of the local region of support employed in feature based methods, there is no discrimination between multiple similar regions. Many a times, the conventional method of matching a feature to the 'best one' is found to be wanting, when applied to document images, due to the presence of multiple occurrence of the same character. To address this problem, we augment the ART coefficients of a CC with those of its 2 nearest neighbors (NN). Then, correspondence is established using these augmented features. We declare 2 CCs as a matched pair if they have a mutual nearest neighbor relationship.

If $\mathbf{X}_i = (x_i, y_i, 1)^T$ and $\mathbf{X}'_i = (x'_i, y'_i, 1)^T$ represent homogeneous coordinates of the matched points in the reference and the target image respectively, they are related as



(a) (b) (c)

Figure 3. Illustration of matching with and without augmented features. The best 9 matches of a component 's' indicated in (a) are shown along with the order of their 'similarity' scores in (b). In the presence of multiple similar regions, the best match is seldom the correct one. However, using the augmented features, the best match, as indicated in (c), is the correct one. The use of augmented features significantly increases the number of correct matches.

follows:

$$\mathbf{X}'_i = \mathbf{H}\mathbf{X}_i \tag{7}$$

where $\mathbf{H} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix}$ is a homography that relates the input images. The equality in Eqn. 7 is defined up to a scale factor. We use the direct linear transform as described in [5] in conjunction with RANSAC algorithm [3] to obtain a set of feature correspondences consistent with a homography. Using this estimated homography, we perform guided feature matching by looking for the best match in the neighborhood of its projected point. We accept the match to be correct if the distance of the best match is less than $(\mu_{NN} + \sigma_{NN})$, where μ_{NN} and σ_{NN} are the mean and standard deviation of the NN distances of the matched pairs. Finally, a least squares estimate of $\mathbf{H}$ is computed using all the consistent matched pairs.

3.4 Warping and Blending

The final step of mosaic generation is to project the input images onto a common coordinate system (chosen as that of the reference image). The pixels in the target image are warped onto the reference frame using backward algorithm. Finally, the registered images are composited using a biquadratic blending function to eliminate intensity discontinuities that occur at the image boundaries.

$$\mathbf{b}(x,y) = \left[1 - \left(\frac{x - x_0}{x_0}\right)^2\right]\left[1 - \left(\frac{y - y_0}{y_0}\right)^2\right] \tag{8}$$

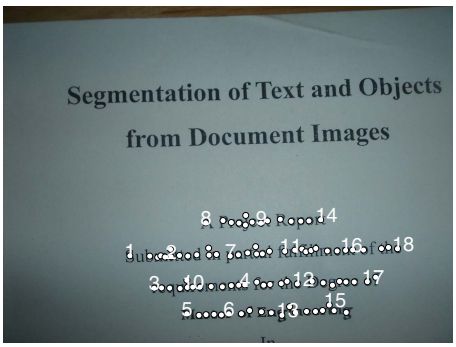
where $1 \leq x \leq \mathrm{M}, 1 \leq y \leq \mathrm{N}$ with M and N denoting the dimensions of the image and (x_0, y_0) is the coordinate of the image center.

4 Experiments and Results

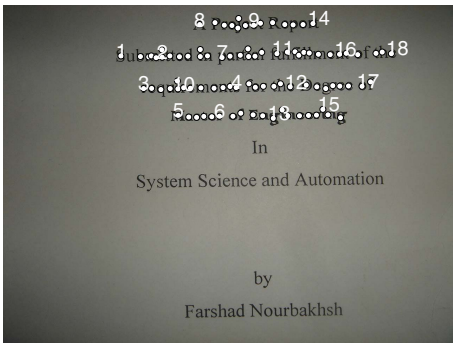
The proposed method is tested on a number of document images acquired using a hand-held camera at a resolution of 1280×960. Performing a CC labelling on the edge image, all the characters in the document, irrespective of the polarity of their foreground and background intensities, are identified for description. We have used the values of 0.2 and 0.3 as thresholds for the hysteresis thresholding step of Canny edge detection. The CCs and their associated measurement regions are identified and normalized to a standard size of 33×33. Feature vectors are matched in pairs across images. A component from the reference image is declared a correct match to one from the target image if they have mutual nearest neighbor relationship.

Fig. 3 illustrates the ineffectiveness of the conventional way of matching a feature to the most similar one in the other image. Fig. 3(b) shows the best 9 matches along with their rank of 'similarity'; all of them correspond to the same character. This situation is highly probable while matching document images. Clearly, choosing the 'best' match would lead to a lot of false matches, that render even robust algorithms like RANSAC ineffective. On the other hand, using the augmented feature, correspondence can be successfully established (See Fig. 3(c)) even when there are several instances of the same letter in the document. The 2-NN constraint, imposed by the augmented features, resolves the ambiguity due to the presence of multiple similar components. This significantly increases the proportion of inliers in the putative matches. Few cases of wrong matches that may arise due to repeated words are effectively handled by RANSAC algorithm.

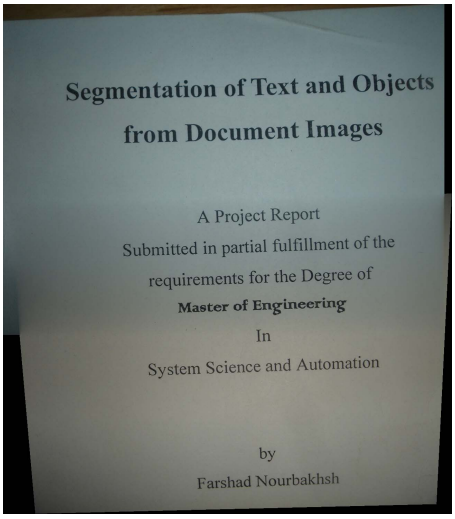
Since CCD is computed for every CC irrespective of its location, it is invariant to translation. In addition, it has all



(a)

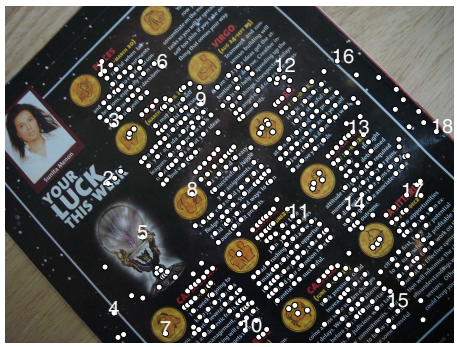


(b)

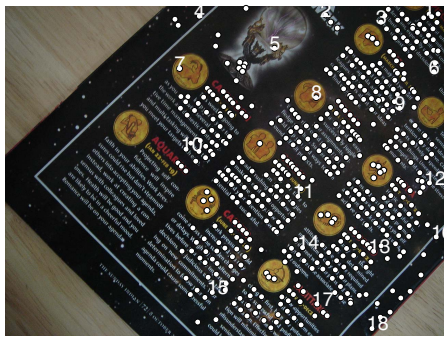


(c)

Figure 4. Result of feature matching across a pair of images (a) and (b) exhibiting translation with different exposure. Of the 94 correctly matched components, only a few of them are indicated for the sake of clarity. The corresponding mosaic is shown in (c).



(a)



(b)

(c)

Figure 5. The matched pairs (521 in number) obtained using CCD for an input image pair (a) and (b) having a large rotation of about 90^{o}. For clarity, only a few of the matched pairs are labeled. The corresponding mosaiced output is shown in (c).

the robust characteristics of ART. Size normalization of the measurement regions effectively handles large changes in scale. Fig. 4 shows the result of feature matching for an image pair exhibiting translation and different exposure. The matched pairs obtained using CCD are overlaid on the respective input images. Fig. 5 demonstrates the rotation inv-

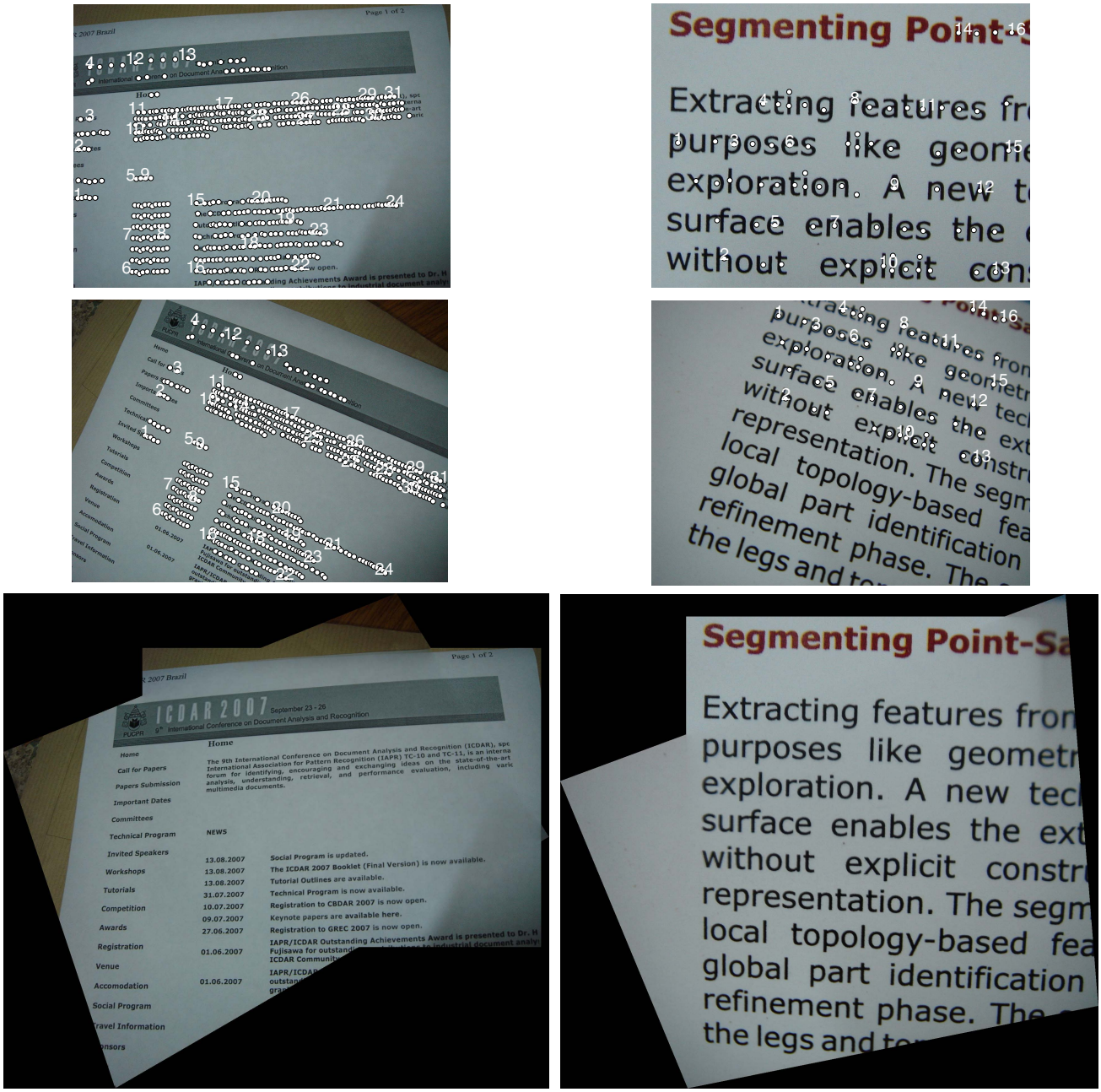


Figure 6. Successful feature correspondence across images with composite rotation, scale and perspective distortion. The matched pairs obtained using CCD are overlaid on the respective input images. Only a few of them are labeled for better visibility. The corresponding mosaiced outputs are shown in the last row.

ariant property of CCD. Feature correspondence is successfully established across images with a large rotation difference of about 90°. In Fig. 6, we consider images taken under general viewing conditions that can have rotation, scale and perspective distortion. A large number of feature matches is obtained, as desired, from the regions common to both the input images. Using the computed feature correspondences, the input images are registered. The final mosaiced output images, obtained after blending, are shown along with the corresponding input image pairs.

5 Conclusions

This paper introduces a new region-based descriptor, well suited for document image processing. Connected component is a natural candidate for localization of features in document images since they are highly stable and largely invariant to geometric and photometric distortions. The proposed method is thus guaranteed to have stable feature localization, which is a critical requirement in all feature-based approaches. The robustness of CCD is amply illustrated by our experiments. The discriminative power of CCD is further enhanced by augmenting it with those of its geometric neighbors. This ensures successful feature correspondence even in the case of occurrence of the same character in multiple locations in the the document. However, the method may fail for images of poor resolution and natural images where connected components cannot be accurately identified.

Since CC labeling is a fundamental processing step in all OCR systems, the method can easily be augmented with subsequent processing modules required for recognition.

References

[1] M. Bober. MPEG-7 visual shape descriptors. *IEEE Trans. Circuits Systems and Video Technology*, 11(6):716–719, 2001.

[2] J. Canny. A computational approach to edge detection. *IEEE Trans. PAMI*, 8(6):679698, 1986.

[3] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commmunications of the AGM*, 24(6):381–395, 1981.

[4] C. Harris and M. Stephens. A combined corner and edge detector. *In Proc. 4th Alvey Vision Conf.*, pages 147–151, 1988.

[5] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2002.

[6] S. K. Hwang and W. Y. Kim. Fast and efficient method for computing ART. *IEEE Trans. Image Processing*, 15(1):112–117, 2006.

[7] T. Kasar, J. Kumar, and A. G. Ramakrishnan. Font and background color independent text binarization. *Proc. Workshop on Camera Based Document Analysis and Recognition*, pages 3–9, 2007.

[8] Y. S. Kim and W. Y. Kim. A new region-based descriptor. *ISO/IEC MPEG99/M5472, Maui, Hawaii*, Dec 1999.

[9] J. Lian, D. DeMenthon, and D. Doermann. Camera-based document image mosaicing. *Proc. Intl. Conf. Pattern Recognition*, 2006.

[10] D. Lowe. Object recognition from local scale-invariant features. *Proc. Intl. Conf. Computer Vision*, pages 1150–1157, 1999.

[11] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2:257–263, 2003.

[12] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Intl. Jl. Computer Vision*, 65:43–72, 2005.

[13] E. N. Mortensen, H. Deng, and L. Shapiro. A SIFT descriptor with global context. *Proc. IEEE conf. Computer Vision and Pattern Recognition*, 1:184–190, 2005.

[14] M. Pilu and F. Isgro. A fast and reliable planar registration method with applications to document stitching. *Proc. IEEE workshop on Application of Computer Vision*, pages 245–250, 2002.

[15] S. M. Smith and J. M. Brady. Susan: A new approach to low-level image processing. *Intl. Jl. Computer Vision*, 23(1):45–78, 1997.

[16] A. P. Wichello and H. Yan. Document image mosaicing. *Proc. Intl. Conf. Pattern Recognition*, 2:1081–1083, 1998.

[17] A. Zappala and a. T. M. Gee, A. Document mosaicing. *Image and Vision Understanding*, 17:589–595, 1999.