# Block-Based Feature Detection and Matching for 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 - 560 012
tkasar, ramkiag@ee.iisc.ernet.in

## Abstract

*In this paper, we present a new feature-based approach for mosaicing of camera-captured document images. A novel block-based scheme is employed to ensure that corners can be reliably detected over a wide range of images. 2-D discrete cosine transform is computed for image blocks defined around each of the detected corners and a small subset of the coefficients is used as a feature vector. A 2-pass feature matching is performed to establish point correspondences from which the homography relating the input images could be computed. The algorithm is tested on a number of complex document images casually taken from a hand-held camera yielding convincing results.*

## 1. Introduction

Mosaics of images have a number of interesting applications such as creating virtual environments [8], panoramic images [7], representing and indexing video information [4]. Document image analysis 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. Traditionally, flatbed scanners and mounted imaging devices are used for document image analysis. However, digital cameras have become increasingly popular as as alternative imaging device. Unlike scanners, they are handy and can capture images of bound books, fragile historical manuscripts and text in scenes, thereby offering much more flexibility to the user. Thus, camera-based image analysis has several potential applications like licence plate recognition, road sign recognition, digital note-taking, document archiving and wearable computing. However, these advantages come at the cost of uneven lighting, low resolution, blur, and perspective distortion. The ability to tackle these challenges will help us effortlessly obtain and manage information in documents.

## 2. Review of Previous Work

Several approaches have been proposed for document image mosaicing. Wichello and Yan [9] 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. Also, it was assumed that there is no warping, which is generally not so in the case of camera-captured images. Zappala et al [10] 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. Recently, Lian et al [5] have proposed a 2-step approach for mosaicing without restricting the motion of the camera. 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. The above methods are suitable for documents which are predominantly text. It may not work well for documents with complex layout and those containing a lot of other entities like pictures and tables. Segmentation of lines, columns and words may not be possible for such complex documents.

We address these issues by employing 'corner' features that can be abundantly detected in text regions as well as in images. Hence the method is applicable to more general documents. In this work, no knowledge of the camera pa-

rameters is assumed and the success of the algorithm solely depends on the features which should be robust to photometric and geometric distortions. The block based method for reliable corner detection and matching is discussed in detail in the following section. The feature correspondence thus accomplished is used to estimate the homography and the subsequent results of mosaicing are presented next. Finally, the conclusions drawn are given in the end.

## 3 Proposed Method

We have chosen a block-based approach for mosaicing of camera-captured document images. Unlike scanned images, camera-captured images generally have uneven lighting and there is very little we can do about it during image acquisition. Corner detection itself may not be reliable under such varying lighting conditions. We have used the Harris corner detector [2] for feature localization. The general procedure of corner detection is to set a fixed threshold on the corner strength and then perform a non-maximal suppression on the corner response to obtain only the locally dominant features. Using a fixed threshold may sometimes result in prohibitively high number of features being detected, especially in document and highly textured images or it may yield too few features if the image has a poor contrast. It is also important to have the corners well distributed over the input images so that a better estimate of the homography could be obtained. We address these issues by employing a new scheme for non-maximal suppression. A 2-pass feature matching is carried out to establish point correspondences from which the homography relating the input images could be computed.

### 3.1. Feature Detection

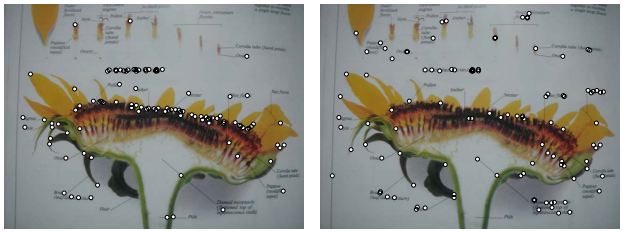
Corners have been the most widely used feature owing to its 2-D structure that provides maximum information content. We have employed the Harris corner detector which is widely used due to its robustness to rotation, translation and changes in view-point and illumination. The Harris corner response $\mathbf{R}$ at the pixel location $(x, y)$ is computed based on the local characteristics of the first order derivatives as follows:

$$\mathbf{R} = \frac{\mathrm{Det}(\mathbf{M})}{\mathrm{Trace}(\mathbf{M})} \tag{1}$$

$$\text{where} \quad \mathbf{M} = \begin{pmatrix} \mathrm{G}_\sigma \otimes \mathrm{I}_x^2 & \mathrm{G}_\sigma \otimes \mathrm{I}_x\mathrm{I}_y \\ \mathrm{G}_\sigma \otimes \mathrm{I}_x\mathrm{I}_y & \mathrm{G}_\sigma \otimes \mathrm{I}_y^2 \end{pmatrix} \tag{2}$$

$\mathrm{G}_\sigma$ is a gaussian function with variance $\sigma$, $\otimes$ denotes convolution operation and $\mathrm{I}_x$ and $\mathrm{I}_y$ are the first order derivatives at the location $(x, y)$ along the x and y directions respectively. Gaussian smoothing avoids corners being detected due to noise. Corners are defined as local maxima of the response function $\mathbf{R}$. The response function $\mathbf{R}$ is analyzed in a block by block manner using two locally adaptive thresholds $\mathrm{T}_1$ and $\mathrm{T}_2$. This results in locating salient features which would not have been detected using a single global threshold for the whole image. But at the same time, detecting corners in each block yields features even in smooth and low-contrast regions. We avoid this by detecting the features only if the maximum of the each block is at least $\mathrm{T}_1\%$ of the global maximum. The result of the approach is that we get 'strong' corners (whose response values are greater than $\mathrm{T}_2$) that are evenly distributed all over the image and enhances the homography estimate. The method is found to consistently yield 'good' corner localization over a number of images and it also does away with the problem of choosing a single fixed threshold. The block based corner detection scheme may be summarized as follows:

1. Compute the maximum $\mathbf{G}_{max}$ of the corner response function values
2. For each block, do
- Compute the maximum response $\mathbf{B}_{max}$ of the block
- If ($\mathbf{B}_{max} > \mathrm{T}_1\%$ of $\mathbf{G}_{max}$ )
  Locate pixels with response value more than $\mathrm{T}_2\%$ of $\mathbf{B}_{max}$
  Perform non-maximal suppression
- Else
  go to the next block
- End



(a) (b)

**Figure 1. (a) Conventional Harris corner detection approach of choosing only the 'strong' corner responses above a fixed threshold as compared to (b) Block-based non-maximal suppression . The new scheme ensures more even spatial distribution of the features while at the same time captures many salient features missed out in conventional method. Employing locally adaptive thresholds alleviates the need to choose a fixed threshold**

### 3.2. DCT Feature descriptor

We have employed a simple yet effective feature matching scheme using the discrete cosine transform (DCT). It has a number of desirable features with good energy compacting property. The feature vector is formed by computing the DCT of the local image region around each of the detected corners. A compact 24-dimensional feature vector is formed by taking the low frequency DCT coefficients to describe the local image region under consideration. The high frequencies which are generally corrupted by image noise are discarded and hence the feature vector is less sensitive to imprecisions that could be present in the image patches.

For a block of image (**U**) of size N×N, its DCT (**V**) is obtained from the following equation:

$$\mathbf{V} = \mathbf{CUC}^T \tag{3}$$

where **C** = $\{c(k,l)\}$ is the DCT transformation matrix defined as follows:

$$\begin{aligned} c(k,l) &= \sqrt{\frac{1}{N}}, \quad k = 0, \quad 0 \le l \le N-1 \\ &= \sqrt{\frac{2}{N}} \frac{\cos(2l+1)\pi k}{2N}, \\ & \quad 1 \le k \le N-1, \quad 0 \le l \le N-1 \end{aligned} \tag{4}$$

Here $k$ and $l$ denote the row and column indices respectively. We consider a local region around each detected corner specified by a square window of fixed size and 2-D DCT is computed for each of these image blocks. We consider the 5×5 sub-matrix on the top-left corner of the DCT coefficient matrix and retain only the 24 AC coefficients, normalized to unit variance, as the feature vector.

### 3.3. Matching and Homography Estimation

Feature matching is one of the most difficult problem. There are inevitably some corners which do not have a match in the other image. There are always some false matches, mainly due to similar and repeated structures in the images. This is in fact the case with document images where the same character may be interspersed throughout the entire page. Lowe [6] suggested an effective measure for feature matching by comparing the distance of the nearest neighbor (NN) to that of the second nearest neighbor. We declare a match between two features to be correct if

$$\frac{\text{1-NN distance}}{\text{2-NN distance}} \le th \tag{5}$$

Here, 'th' is a threshold set to 0.8 in our work. This measure performs well because of the fact that correct matches have the nearest neighbor significantly closer than the closest incorrect match. False matches as well as ambiguity in the case of multiple matches are thus effectively eliminated. The feature matches thus obtained are refined using RANSAC [1] to obtain a homography consistent with a large number of matched features as described in [3].

Finally, we compute the rectangular bounding box of the set of matched features which roughly represents the region of overlap between the images under consideration. Another pass of the feature localization and matching is performed again within this region of overlap to get a denser point correspondence.

### 3.4. Warping and Blending

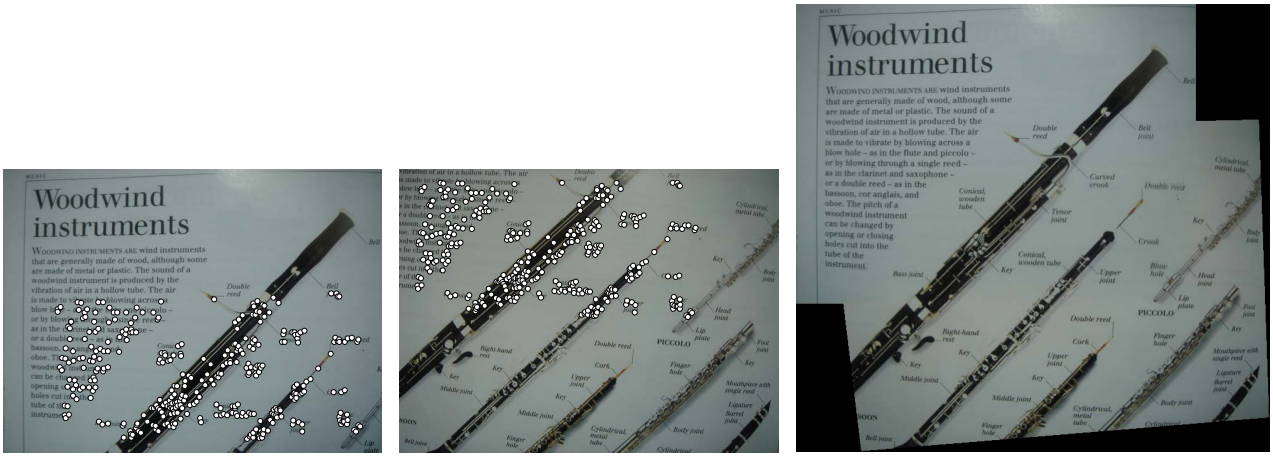
The matches obtained as explained above are then used to compute a least squares estimate of the homography. The target image is projected onto the plane of the reference image so that the two images share a common coordinate system. Furthermore, a simple feathered blending eliminates visible discontinuities at the image boundaries that remain after warping the target image onto the reference image. The blending function **B** assigns highest weight to the pixel at the image center and the weights gradually decrease in both the image dimensions towards the boundary.

$$\mathbf{B}(x,y) = (1 - (\frac{x - x_0}{x_0})^2)(1 - (\frac{y - y_0}{y_0})^2) \tag{6}$$

where $1 \le x \le M, 1 \le y \le N$ with $M$ and $N$ representing the dimensions of the image and $(x_0, y_0)$ is the image center. This results in a seamless mosaic image.

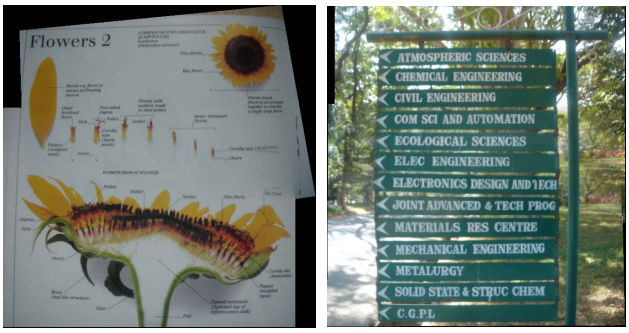
## 4 Experimental Results

The test images were acquired from a hand-held camera at a resolution of 1280×960. The above algorithm is tested on a number of challenging document images. The image is subdivided into 64 equal blocks to select the Harris corners. The parameters used for the Harris corner detection are $\sigma = 2$ and 3×3 window for non-maximal suppression. The threshold criteria for non-maximal suppression are set at 70% of the maxima of each individual block which in turn is constrained to be at least 1% of the global maxima. Around each detected corners, a window of size 31×31 is considered and this local image region is described by its 2-D DCT coefficients. The feature matching scheme using DCT gives a good performance with a large proportion of correct matches thereby facilitating a fast homography estimation via RANSAC. Fig. 2 shows the results of feature matching. The mosaics of different types of input images obtained after blending are shown in Fig. 3.




(a) (b) (c)

**Figure 2. The result of feature matching (375 matches) are shown in (a) and (b) while the corresponding mosaic is shown in (c)**




(a) (b)

**Figure 3. Mosaic outputs of documents with (a) complex layout and (b) scene text**

## 5 Conclusions and Future Work

We have presented a simple yet effective scheme for feature matching across uncalibrated camera captured document images. The Harris corners have been meticulously selected so that we have the features fairly evenly spread throughout the image. The problem of extracting prohibitively high or too few number of features are effectively by-passed by employing two pseudo-adaptive thresholds. The features are described by the DCT coefficients evaluated around each detected corner. The ease of obtaining the features and its low dimensionality is definitely an edge over other existing methods. Despite its simplicity and low dimensionality, the DCT features have been found to have high discriminating power. It works well for wide range of complex document images with variations in illumination and viewpoint changes. But, it is sensitive to scale and large rotations and this deserves further study. Page curl correction in the case of bound volumes and removal of perspective distortion need to be incorporated. Automated text localization and recognition from camera images is our future work.

## References

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

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

[3] R. Hartley and A. Zisserman. *Multiple view Geometry in Computer Vision*. Cambridge University Press, 2003.

[4] M. Irani and P. Anandan. Video indexing basedc on mosaic representation. *Proc. IEEE*, 86(5):905–921, 1998.

[5] J. Lian, D. DeMenthon, and D. Doermann. Camera-based document image mosaicing. *ICPR*, 2006.

[6] D. G. Lowe. Object recognition from local scale-invariant features. *ICCV*, pages 682–688, 1999.

[7] S. Peleg and J. Herman. Paranomic mosaicing with videobrush. *DARPA Image Understanding Workshop*, pages 261–264, 1997.

[8] R. Szeliski. Video mosaics for virtual environments. *IEEE computer Graph Appl.*, 16:22–30, 1996.

[9] A. P. Wichello and H. Yan. Document image mosaicing. *ICPR*, 2:1081–1083, 1998.

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