Multi-Resolution Probabilistic Information Fusion for Camera-based Document Image Matching

Sumantra Dutta Roy*, Sitanshu Gupta†, Ishaan Gupta‡, Kavita Bhardwaj*, Santanu Chaudhury*
*Dept of EE, IIT Delhi †Dept of ECE, MNNIT ‡Dept of ECE, NSIT
New Delhi - 110016, INDIA Allahabad, INDIA New Delhi - 110078, INDIA
Email: sumantra@ee.iitd.ac.in, sitanshu.08@gmail.com, ishaan09@gmail.com,
kavitabhardwaj.iitd@gmail.com, schaudhury@gmail.com

***Abstract*—Given a part of a document image taken with any camera at an arbitrary orientation and sometimes far-from-perfect illumination, an important problem is to match this query image to the corresponding full image from a document database. We propose a novel multi-resolution robust methodology for the same. The method combines information from independent sources of measurement in a probabilistic framework. The proposed method is invariant to a large range of distortions, illumination changes, and is relatively resilient to noise and unmodelled objects present as clutter. To the best of our knowledge, no related work address all these issues.**

Keywords:

Multi-resolution Analysis, Projective Transformation, Homography, Robust Error Norm, Contour Envelope of Text Block, Geometric Hashing, Probabilistic Information Fusion

I. Introduction

We present a novel multi-resolution probabilistic method for matching a database document to a degraded query image (for instance, taken from a low quality camera in bad illumination and even with a part of the document occluded with other objects.) The ubiquitous nature of mobile phones with cameras makes camera-based document image analysis an important research [1], [2]. An interesting development in public image collections such as Flickr is the presence of images stored at multiple resolutions. Our work explores the speed-resolution trade-off to propose an efficient matching strategy. Our system does not consider database organisation issues which affect efficiency in image retrieval: we present a novel approach to the image matching problem.

Document image retrieval includes problems such as queries about layout [3] and logos [4]. These methods do not suffice for the problem where a partial snapshot of text part of a document is the query image. Nakai et al. [1] consider the problem of camera-based document image retrieval. The authors model distortions using projective invariants (cross ratios). They consider features as centroids of word regions, and use cross ratios to vote for documents in a Hough transform-like manner, with a few problem-specific heuristics. The number of feature points is often too large, and the voting-based procedure takes too much running time. In [2] Liu and Doermann propose an approach in which layout context dictates local features. Since this relies on the layout of words, it fails when there is a small amount of text present in captured image. It also approximates the projective transform locally, so it is not invariant for significant perspective distortions. The Kise group extend their earlier ideas in [1] and experiment with affine and projective models. In [5] Nakai et al. propose the approach where combination of local invariants are hashed into large hash-table. The hash table gives the process polynomial time complexity. In Nakai and Kise [6], the authors use their Locally Likely Arrangement Hashing (LLAH) with affine invariants, using a neighbourhood assumption. This simplifies the polynomial time complexity of a geometric hashing-based strategy to a linear one. We propose an alternate strategy based on a multi-resolution approach. We use more than one feature, for robustness. One feature is similar to that of the Kise group (contour envelope curvature extrema as opposed to their word centroids) with geometric hashing: on an average, we deal with far fewer points than them. Further, a multi-resolution approach in which we often operate at a very low scale - reduce our computational loads immensely. We could apply a LLAH-like strategy to reduce this load even further. Existing algorithms do not consider images with noisy elements such as a pen, hand, or small objects, as shown in Fig 4. Our proposed technique is entirely script- and language-independent: it does not use any features of any specific script, or language e.g., Fig. 3. To the best of our knowledge, no relevant work addresses all these issues.

II. A Robust Multi-Resolution Approach with Multiple Independent Sources of Measurement

In this section, we first examine the wide variety in query images that can be submitted to the system (Sec. II-A). Sec. II-B considers fusion of probability estimates from multiple independent sources of measurement. Sections II-C and II-D consider the two features used in this work namely, text/image blocks, and the extrema points of contour envelopes, and discuss issues related to handling these features at multiple levels of resolution. Sec. II-E explains the pre-processing steps in order to extract these two features from a given document image.

A. Wide Variations in Query Images: Geometric Deformations, Illumination Variations, Noise

Database images are generally taken in good imaging conditions with good and uniform illumination and zero skew. For a query image, a common situation is to have a part of a document image taken by a common camera (a cellphone camera, for instance), and at an arbitrary orientation, and possibly in a region of bad illumination. Further, there could be structured and/or unstructured noise in the image: imaging noise, or other objects occluding parts of the document.

In general, the geometric deformation could be non-linear. We use general linear model to approximate the deformation: A 2-D projective transformation. The fundamental theorem of Plane Projective Geometry (extensively cited in [7]) relates any two planes in higher dimensional space using a 2-D projective transform. Hence, the features used for matching have to be either projective invariant, or estimating the homography between two projective planes.

To encounter the effects of illumination variation in the query image, we have a relative gradient image [8]:

$$I(x,y) = \frac{|\nabla F(x,y)|}{\max_{(u,v)\in W(x,y)} |\nabla F(u,v)| + c} \tag{1}$$

where I is the relative image gradient. In this equation, $F(x,y)$ denotes the image intensity, the ∇ denotes the intensity gradient, $W(x,y)$ is a local window centred at pixel (x,y) and c is a small positive constant used to avoid division by zero. We replace all expressions involving pixel intensities with the relative gradient, mentioned above.

B. Multiple Sources of Measurement

The proposed technique is independent of the specific sources of measurements for different features. The database of documents $\mathcal{D}$ contains m documents D_1, D_2 $\ldots D_m$. Each document D_i has text/image blocks d_{ik_j}. Consider a query image Q with n text/image blocks q_1, q_2 $\ldots q_n$. Consider block q_j in the query image. This could correspond to any block d_{ik_j}. Let $P_{f_l}(q_j|d_{ik_j})$ denote the probability of query image block q_j corresponding to block d_{ik_j} in document D_i, obtained using feature f_l. Using the features f_l (which come from independent sources of measurement), we define the total probability of the query image block q_j being block d_{ik_j} as

$$P(q_j|d_{ik_j}) = \prod_l P_{f_l}(q_j|d_{ik_j}) \tag{2}$$

For our experiments, we use two features: the bounding quadrilateral around the text/image block (Sec. II-C) and the block contour envelope curvature extrema projective co-ordinates (Sec. II-D). Sections II-C and II-D describe the computation of the corresponding $P_{f_l}(q_j|d_{ik_j})$ for the two cases, respectively.

C. Script-Independent Matching of Text/Image Blocks

The first feature that we use are four corner points of the bounding quadrilateral of a text or an image block. (Section II-E outlines the basic pre-processing steps in our system). A block q_j in the query image Q could correspond to a database block d_{ik_j} of document D_i. We model the probability of the block in question being d_{ik_j} given that we have observed query image block q_j, as follows:

$$P_{f_l}(d_{ik_j}|q_j) = 1 - (1/R)\sum_r \rho(x_r, \sigma_1) \tag{3}$$

Here, $\rho(x,\sigma)$ denotes the robust error norm [8], where σ is a scale factor:

$$\rho(x,\sigma) = \frac{x^2}{x^2+\sigma^2} \tag{4}$$

The above summation is for all pixels r in the warped query block, with respect to the corresponding pixels in the database document block d_{ik_j}, and R is the total number of such pixels. Let x denote the pixel intensity difference (based on the relative gradient: Sec. II-A) between corresponding pixels of the projected query block, and a database document block, for a given pixel location. The advantage of taking $\rho(x,\sigma)$ in place of x^2 (or a normalised version of it, for that matter) is that the robust error norm is more robust to an outlier. Corresponding to an outlier pixel, if $|x| > \sigma/3$, its influence on the solution will be less as $\rho(\cdot)$ approaches 1.

1) Selecting the Right Scale: The system starts at the smallest resolution. Database images are stored at different resolutions. The system uses the information from the smearing algorithm (Sec. II-E) to obtain an estimate of the font size (we assume that documents will have at least some text in them.)

D. Geometric Hashing-based Matching of Contour Envelope Curvature Extrema Projective Co-ordinates

From the basic pre-processing steps of Sec. II-E, the second feature we use is the curvature extrema of the contour envelope. For an image at any level in the Gaussian pyramid, smearing results in a text block. For each such block, we use the standard parametric representation for curvature:

$$\kappa = \frac{|x'y'' - x''y'|}{(x'^2 + y'^2)^{\frac{3}{2}}} \tag{5}$$

Here, the y co-ordinate and the x coordinate of every pixel is assumed to be a function of the index number of the point on the contour and the derivatives (y', x', y'', x'') are accordingly calculated using approximating difference equations.

Consider a text block q_j from a query image Q containing N_j curvature extrema points. To consider the match with block d_{ik_j} with M_j curvature extrema points, we note that a naive strategy to match N_j points with M_j points would incur exponential time complexity. To reduce this to polynomial time, we use a Geometric Hashing-based strategy. We consider a hash table for both the database document block d_{ik_j},as well as the query block q_j. We can select ordered set of 4 basis points from the database block in $\binom{M_j}{4} \times 4!$ ways - this is $\mathcal{O}(M^4)$. (We can reduce this by not considering all 4! combinations, since realistic imagining conditions preclude all but 4 of these [7]). For every quartet of basis points selected, a *Hash Table* stores the projective coordinates of the rest of

the $M_j - 4$ curvature extrema points. We perform the same procedure for a query image block.

Block matching between a database document and query block reduces to matching rows of two hash tables. Matching two table rows has quadratic time complexity. We can reduce this to linear, if each has table row is sorted. Hence, the problem of matching curvature extrema reduces to $\mathcal{O}(M_j^5 N_j^5)\times$ *the row matching time*.

1) Selecting the Right Scale: Just as the resolution determines the number of pixels in a block (Sec: II-C1), it determines the number of contour extrema in a curve (contour) represented at different resolutions/scales.

E. Feature Extraction

Both features in Sections II-C and II-D have a common processing pipeline. Images are stored at multiple levels of resolution. The first step is the application of a run length smearing algorithm [9]. We learn these smearing parameters (such as the horizontal and vertical run-lengths) for a large number of documents at different scales, and store them in a look-up table. For an image at a given scale, we use a simple sequential labelling-based segmentation algorithm to find the number of connected regions (blocks).

For the first feature, we use a Hough Transform-based method to fit a quadrilateral around a text/image block provided it is greater than a particular size (this is again a scale-dependent parameter). We do this only for blocks for which it is possible to fit four lines around it. For the second feature (curvature extrema points), we operate straight on the contour of the extracted block.

III. Probabilistic Hypothesis Generation

Given a database of m documents, we initialise $P(D_i) = 1/m$, where $P(D_i)$ is the *a priori* probability of document D_i. We are given a query image Q (which contains n blocks q_j). A query block q_j could correspond to a database document block d_{ik_j} in document D_i. Based on the features in Sections II-C and II-D, we compute the probability that a particular query image block q_j corresponds to database document block d_{ik_j} as follows. For a system with l features f_1, f_2 $\ldots f_l$, we compute this probability as

$$P(d_{ik_j}|q_j) = \prod_l P_{f_l}(d_{ik_j}|q_j) \tag{6}$$

This is reasonable, since we assume that the l features and their measurement processes are independent. In our case, we have considered $l = 2$ for the number of features.

Given n blocks detected in the query image Q, the system forms hypothesis corresponding to the correct identity of each query block q_j, $1 \leq j \leq n$. *We note that while all query image blocks have to correspond to one document D_i, one may have more than one hypothesis corresponding to a document D_i.* Given that we have observed query image blocks $q_1 \ldots q_n$, we compute the probability that these n blocks correspond to blocks $d_{i_1} \ldots d_{i_n}$ corresponding to document image D_i:

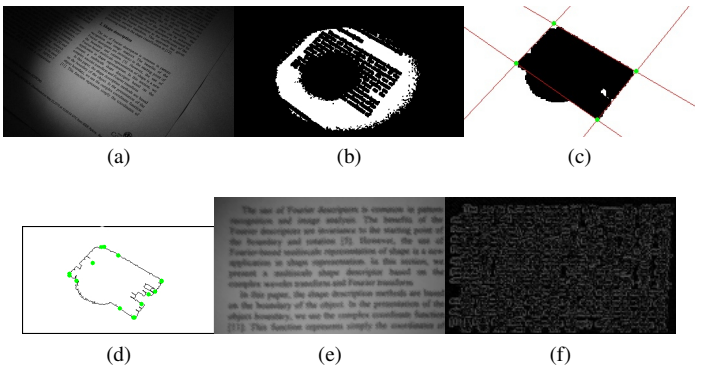


Fig. 1: The system works in cases of bad illumination conditions: an illustration. (a) The original query image, (b) the binarised image, (c) The quadrilateral block feature, (d) the text countour and its curvature extrema, (e) the homography-transformed query block, and (f) the error image: difference between the database document block, and the projected query image block

$$P(d_{i_1} \ldots d_{i_n}|q_1, \ldots q_n) = \frac{P(q_1 \ldots q_n|d_{i_1} \ldots d_{i_n})P(d_{i_1} \ldots d_{i_n})}{\sum_t P(q_1 \ldots q_n|d_{t_1} \ldots d_{t_n})P(d_{t_1} \ldots d_{t_n})} \tag{7}$$

The summation in the denominator is for all hypothesis t corresponding to the identity of all n blocks q_1, q_2 $\ldots q_n$. Further, the second term on the right side of the above equation can be further simplified as follows

$$P(d_{i_1}, d_{i_2} \ldots d_{i_n}) = P(d_{i_1}, d_{i_2} \ldots d_{i_n}|D_i)P(D_i) \tag{8}$$

Here, $P(D_i)$ is the *a priori* probability of document D_i, and the other terms may be approximated by the relative areas of the document blocks in the corresponding database image, in the corresponding same orientation. We compute the first term of the numerator as follows:

$$P(q_1 \ldots q_n|d_{i_1} \ldots d_{i_n}) = \prod_n P(q_j|d_{ik_j}) \tag{9}$$

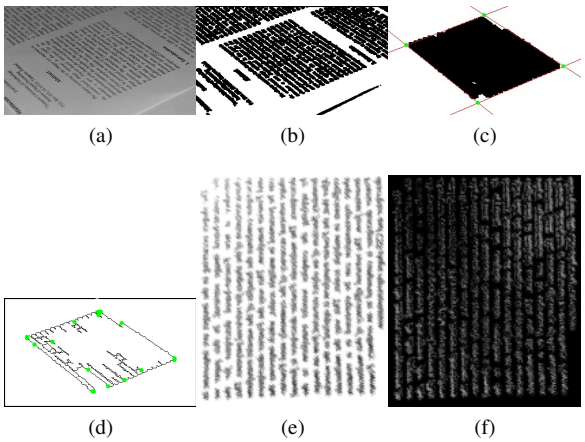
We compute the final *a posteriori* probability of document D_i as the sum of probabilities of all individual hypothesis corresponding to the particular document D_i.

$$P(D_i) = \sum_t P(d_{t_1}, d_{t_2} \ldots d_{t_n}|q_1 \ldots q_n) \tag{10}$$

IV. Experimental Results and Discussion

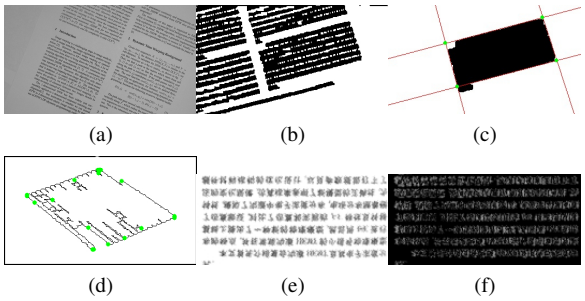
This paper represents work in progress. We have a set of 50 database document images, and 100 query images. The database document images have maximum size at the highest resolution level of 2340×1700, and the corresponding maximum figure for query images is 2848×1600.

1) Experiments with large illumination variations: Fig. 1 shows an example of successful matching in spite of bad illumination conditions.



(a) (b) (c) (d) (e) (f)

Fig. 2: Figure showing the performance of the proposed technique on a highly skewed query image. The different parts of the figure correspond to those in Fig. 1.



(a) (b) (c) (d) (e) (f)

Fig. 3: Script and language independence: successful matching of a Chinese language document. The different sub-parts are same as in Fig. 1.

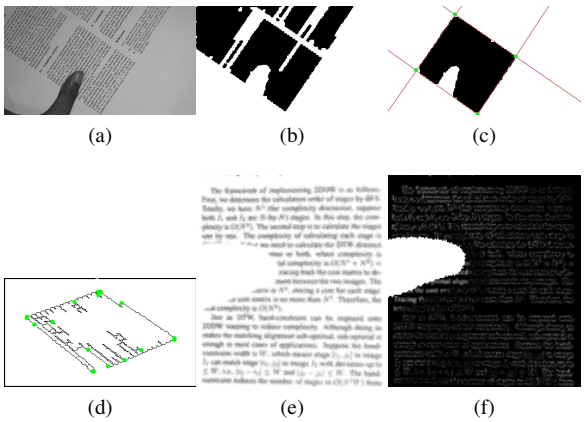
2) Highly skewed query image: Fig. 2 shows an example of successful matching in spite of a large amount of skew in the query image.

3) Script and language independence: An advantage of our system is that it is independent of the specific language/script used in a document image. Fig. 3 shows an example of successful matching on a Chinese language document page.

4) Cases of occlusions, structured noise: Fig. 4 shows an example of successful matching in spite of structured noise: in this case, a common occurrence for images of documents taken with a cellphone camera, or a hand-held camera.

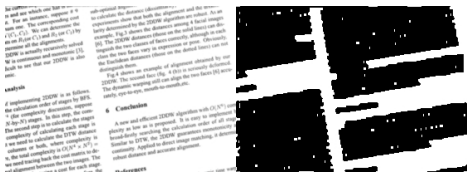
5) Miscellaneous failure cases: Out of 70 query images the system gave correct results (matched the corresponding database document) in 65 cases. The 5 failure cases were due to errors at the block building and feature detection stage itself. Fig. 5 shows such a case.

Some statistics for the above cases are as follows. For 5 out of 10 query images with insufficient illumination, it was not possible to separate the blocks from the background. For the 20 images with occlusions and structured noise, there were 7 failures either because the object was at the corner of a block



(a) (b) (c) (d) (e) (f)

Fig. 4: Successful matching in spite of structured noise: a common case of some text hidden by a finger: a common occurrence for hand-held document images with a cellphone or a camera.



(a) (b)

Fig. 5: A failure case due to errors at the block building and feature detection stage itself.

(resulting in a wrong bounding quadrilateral), or resulted in more contour curvature extrema from the occluding object than from the actual text block.

REFERENCES

[1] T. Nakai, K. Kise, and M. Iwamura, "Camera-Based Document Image Retrieval as Voting for Partial Signatures of Projective Invariants," in *ICS*, 2005.

[2] X. Liu and D. Doermann, "Mobile retriever-Finding the Document with a snapshot," in *Int. Workshop on Camera-Based Document Analysis and Recognition*, 2007, pp. 29–34.

[3] P. Hermann and G. Schlageter, "Retrieval of Document images using layout knowledge," in *Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on*, 1993, pp. 537 – 540.

[4] D. Doermann, E. Rivlin, and I. Weiss, "Applying algebraic and differential invariants for logo recognition," *Mach. Vision Appl.*, vol. 9, pp. 73–86, 1996.

[5] T. Nakai, K. Kise, and M. Iwamura, "Hashing with Local Combinations of Feature Points and its Application to Camera-based Document Image Retrieval," *Proc. CBDAR05*, pp. 87–94, 2005.

[6] ——, "Use of Affine invariants in Locally likely Arrangement Hashing for Camera-based Document Image Retrieval," *Document Analysis Systems VII*, pp. 541–552, 2006.

[7] C. Rothwell, "Recognition using Projective Invariance," Ph.D. dissertation, University of Oxford, 1993.

[8] S.-D. Wei and S.-H. Lai, "Robust and Efficient Image Alignment Based on Relative Gradient Matching," *IP*, vol. 15, no. 10, pp. 2936–2943, 2006.

[9] H. Cao, R. Prasad, P. Natarajan, and E. MacRostie, "Robust Page Segmentation Based on Smearing and Error Correction Unifying Top-Down and Bottom-Up Approaches," in *ICDAR*, 2007.