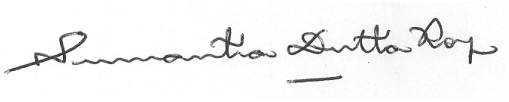
Pattern Recognition Letters

Authorship Confirmation

Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.

As corresponding author I, Sumantra Dutta Roy hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature  Date 01 July, 2014

List any pre-prints:

None.

Relevant Conference publication(s) (submitted, accepted, or published):

None.

Justification for re-publication:

Not applicable

Pattern Recognition Letters
journal homepage: www.elsevier.com

Camera-based document image matching using multi-feature probabilistic information fusion

Sumantra Dutta Roy[a,**], Kavita Bhardwaj[a], Rhishabh Garg[a], Santanu Chaudhury[a]

[a] *Department of Electrical Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi - 110 016, INDIA.*

ARTICLE INFO

Article history:
Received xx xxx 2014
Received in final form xx xxx 2014
Accepted xx xxx 2014
Available online xx xxx 2014

Communicated by S. Sarkar

2000 MSC:
41A05
41A10
65D05
65D17

Keywords:
Camera-based document analysis and retrieval
probabilistic information fusion
Geometric hashing-based matching

ABSTRACT

A common requirement in camera-based document matching and retrieval systems, is to retrieve a document whose image has been taken under difficult imaging conditions (insufficient and non-uniform illumination, skew, occlusions, all of these possibly coming in together in the same image). We present a system for robust matching and retrieval which works well for such difficult query images, using probabilistic information fusion from multiple independent sources of measurement. Our experiments with two robust and computationally inexpensive features show promising results on a representative database, compared with the state-of-the-art in the area.

1. Introduction

The ubiquitous nature of mobile phones necessitates camera-based document image analysis as an important area of research. In the absence of flat-bed scanners and better document imaging devices with proper illumination conditions, a common requirement is to take a quick image of the document with a mobile phone, transmit feature information (or in the worst case, the whole image itself) to a server with access to a database of documents, and retrieve the relevant document. It is common to have a degraded image of part of a document taken by a mobile phone camera in bad, non-uniform illumination, and with a part of it occluded by other objects as well. Fig. 1 shows a few such cases. An important task is to match it with images of documents in a database, without the need for recognition of either the script in the document, or its content.

Fig. 1. A sample input to a document matching system could be taken with a low-quality camera in non-uniform improper illumination, with severe skew and possible occlusion as well. A script-independent system should be able to match the given query image with the corresponding image in the document database.

In this paper, we present a novel multi-feature probabilistic information fusion technique for such a matching task.

Camera-based document image analysis has become an important area of research with the proliferation of mobile phones with cameras (Liang et al. (2005), Liu and Doermann. (2007)). Liang et al. (2005) present a survey of technical challenges

**Corresponding author: Tel.: +91-11-2659-6167; fax: +91-11-2658-1606;
e-mail: sumantra@ee.iitd.ac.in (Sumantra Dutta Roy)

and some solutions for camera-based document images. Hull (1994) proposes a method for organising a database of document images, using image data obtained from the text portion of document images. Liu et al. (2005) propose a method of combining global and local information by using foreground density distribution features and key block features, for document retrieval. Hermann and Schlageter (1993) consider layout information for document retrieval. However, these methods are not applicable to camera-captured documents, as most of the above techniques are for flat-bed scanner outputs.

The state of the art in camera-based retrieval perhaps comes from the *Osaka Prefecture University Group* where Nakai et al. (2005a) extend their earlier ideas and experiment with affine and projective models. Nakai et al. (2005b) propose a combination of local invariants with hashing. Nakai et al. (2006) use their Locally Likely Arrangement Hashing (LLAH) with affine invariants, using a neighbourhood assumption. The most recent work is a highly memory optimised version (Takeda et al. (2011)), which is perhaps the *de facto* yardstick for a camera-based document retrieval system. A recent work (Moraleda (2012)) is reported to do away with a requirement of the above, which needs query images that cover a fair part of the given document page. Moraleda's system works with small patches of blocks in query images (which may be even quite defocussed). Feature vectors are computed using descriptors defined by straight segments connecting word bounding boxes.

We propose a robust matching strategy based on probabilistic fusion of information from multiple independent features. While our theory is independent of specific characteristics of any particular feature, we have experimented with two features, contour extrema (Sec. 2) and zig-zag features (Sec. 3). The first feature relates to the shape of a text block, and the second, on the arrangement of salient parts inside such a block: complementary features, with independent sources of measurement. For the contour envelope feature itself, on an average, we deal with a smaller number of feature points for matching and indexing, as compared to the Osaka Prefecture University systems. We show experimental results of successful matching and retrieval in challenging query document cases, with multiple deformations such as skew, imperfect and non-uniform illumination, and part of the document occluded by fingers or other external objects. To the best of our knowledge, no related work address all these issues.

The layout of the rest of the paper is as follows. Sec. 2 describes the contour extrema point feature in detail. This section presents a probabilistic matching (and subsequent efficient retrieval process) based on geometric hashing. Sec. 3 considers a similar formulation for the second feature (the zig-zag feature), encoding the the relative arrangement of prominent words inside a text block. Sec. 4 proposes a probabilistic fusion of information from various independent sources of measurements, to work in cases where a particular feature does not perform that well due to characteristics of the two feature detection processes, and the noise (random, or structured) affecting the detection of a particular feature. Sec. 5 shows results of successful matching for challenging cases of bad and non-uniform illumination, skew and occlusions, on our representative database. We compare our system with the state-of-the-art in Sec. 6. Sec. 7 concludes the paper.

2. A Geometric Hashing-based Matching Strategy with Contour Extrema Points as Features

We use a multi-scale smearing approach to find text and image blocks in a given image Q, compute its bounding contour, and smooth it with a Gaussian filter. From this, we extract contour curvature local extrema. We consider this as our first feature. In Sec. 3, we consider another feature, and consider the probabilistic combination of independent features, in Sec. 4.

We start with the basic LLAH (Locally Likely Arrangement Hashing) philosophy of the Osaka Prefecture University group, in a slightly different context. Instead of the explicit cross ratio and affine adaptation of the same in the Osaka Prefecture University group (Nakai et al. (2005a)-Takeda et al. (2011)), we use the mathematically equivalent, but more general formulation of geometric hashing (Lamdan and Wolfson (1988)) for a 2-D space with b non-collinear points as a basis ($b = 3$ for an affine space, and 4, for a projective one). This allows us to treat affine and projective invariants in a similar manner, with simply the number of basis vectors varying for the particular invariant considered, instead of having a system-specific pattern of points, as in the work of the above group.

For b non-collinear basis points and M given points, we can construct a hash table with a basis chosen in $\binom{M}{b} \times b!$ ways. For each basis, we can compute the coordinates of the remaining $M - b$ points, leading to a table of size $O(M^{b+1})$. For a projective basis for instance, we have a table of size $O(M^5)$, with each row consisting of projective invariant coordinates $\langle \alpha, \beta \rangle$ pairs. In general, a transformation is rarely linear (the most general 2-D linear transform is a projective one), we assume that the transformation is *locally* linear. We also wish to a fixed size feature vector for each block to aid in the actual hash-based indexing. Hence, we consider only r feature points (the top r curvature local extrema points), and instead of considering all $M - b$ points, we consider only $s - 1$ neighbouring points for each of the r feature points, with these points involved in selecting the b basis points. In our experiments, we have considered a projective basis (the most general linear transform), and empirically considered $r = 30$ and $s = 10$. Another difference between the approach of the Osaka Prefecture University and ours, is that they use word centroids (as an aside, they themselves are only affine invariant, not projective), which are on an average much larger in number than our first feature: high curvature contour points. We discuss other important differences between the two approaches in the discussion section, Sec. 6.

2.1. Probabilistic Matching and Retrieval

Consider a query image Q (such as one of the images in Fig. 1), with n text blocks $q_1, q_2 \ldots q_n$. The database of documents $\mathcal{D}$ contains m documents $D_1, D_2 \ldots D_m$. All the documents D_i in the database are scanned, binarised images with zero skew, with good and uniform illumination. A document D_i has text blocks d_{ik_j}. Consider a block q_j in the query image Q. This could correspond to block d_{ik_j} of database document D_i.

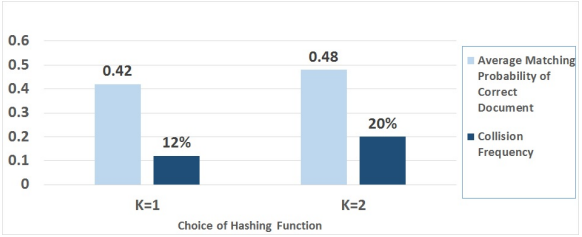


Fig. 2. Experiments with hash function-based retrieval, for Eq. 2, showing the average matching probability of the correct document, and the collision frequency for two choices of the power k, for the first feature (contour extrema): details in Sec. 2.1.

For a given block q_j in a query image Q, we find the distance between the projective invariants. If this distance is less than a threshold, we vote for this pair. For a row of the query image's hash table, the number of votes corresponds to the number of invariants satisfying the condition, above. We find the probability that the query block q_j corresponds to the k_jth block of database document D_i as

$$P(d_{ik_j} | \, q_j) = \frac{\text{votes cast for } d_{ik_j} \text{ by } q_j}{\text{total votes cast by } q_j} \tag{1}$$

Using a method similar to the state-of-the-art retrieval of the Osaka Prefecture University group (Nakai et al. (2005a)-Takeda et al. (2011)), we perform efficient retrieval with the following hash function, using hashing with chaining. The feature vector consists of the $s-4$ projective invariant pairs $\langle \alpha_i, \beta_i \rangle$:

$$H_{index} = \sum_{i=1}^{s-4} (\alpha_i^k + \beta_i^k) mod \; H_{size} \tag{2}$$

where H_{size} is the size of the hash table, and k is a constant that we have empirically considered in our experiments as 1 or 2 (Fig. 2). The probability calculations (as in Eq. 1) are done for all blocks in the chain, which map onto the same index in the hash table.

3. Incorporating Another Feature for Robustness

The matching and retrieval performance of the system described in the previous section (Sec. 2) degrades considerably for cases of severe occlusion, such as the situation shown in Fig. 3, where due to the presence of the occluding hand, a considerable portion of the distorted contour contains a large number of high curvature points. These new points in the occluded region tend to add to false votes, and pull down the number of correct votes. This is especially true for query images having a single block visible. The motivation for using a separate independent feature comes from the fact that even without cases of occlusion, blocks could be similar or mirror images of each other, but could have entirely different content. We need a feature that examines the internal structure of a block as well.

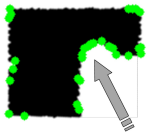


Fig. 3. Occlusion (a white gloved finger on top of a text block, distorting the distribution of contour extrema information, This necessitates the use of another feature, one that relies on the content inside a block: Sec. 3.

A recent work published in the same journal (Moraleda (2012)), describes a robust image matching system for a given blurry query image of a small patch of a document. *We emphasise that while the method is not strictly based on projective invariance (it uses Euclidean parameters, which can handle only limited projective distortions), it is shown to work rather well for limited projective distortions in query images.* The feature considers word bounding boxes and their relative arrangement in a text block. We adapt this feature to our probabilistic matching strategy, based on geometric hashing.

This feature considers bounding boxes around words which are not too small in size. Moraleda bases this on the hypothesis that these contain more relevant information anyway, about the text in the document. Moraleda defines a *zig-zag feature vector* as follows. Fig. 4 shows an example of a 4-segment zig-zag feature vector, and its quantisation. For a 4-segment pattern, Moraleda's system considers a set of four angle pairs $\{(\theta_i, \phi_i)\}$ encoded as a quantised string, which he calls a 'Synthetic word', with each quantised direction encoded as an alphanumeric symbol. While Moraleda uses the 'synthetic word' string as an input to an inverted index search, we adapt his technique to our problem of probabilistic matching based on geometric hashing. As in Moraleda (2012), we further convert the alphanumeric symbols of the 'synthetic word' into a numeric entity, and have a probability-based function as in Eq. 1, with the difference being the angle-based feature vector in this case replacing the invariant-based feature vector, for the first feature. As for the previous feature, we use a hash function:

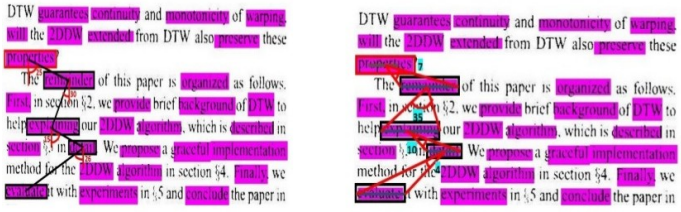


Fig. 4. The Zig-zag feature vector using word bounding boxes, yields a feature vector from the four segment pattern as [25, 7, 30, 35, 15, 10, 26, 4]: Details in Sec. 3.

$$H_{index} = \sum_{i=1}^{k} (\theta_i^u + \phi_i^u) \; mod \; H_{size} \tag{3}$$

where H_{size} is the size of the hash table. Fig. 5 justifies our choice of $k = 4$ segments for the second (zig-zag) feature. As before, the probability calculations are performed only for the blocks in the chaining list: all those with collisions at the same index value.

4. Multi-feature Probabilistic Information Fusion

We are given a query image Q containing n text blocks q_j, $j = 1 \ldots n$. A query block q_j could correspond to the database

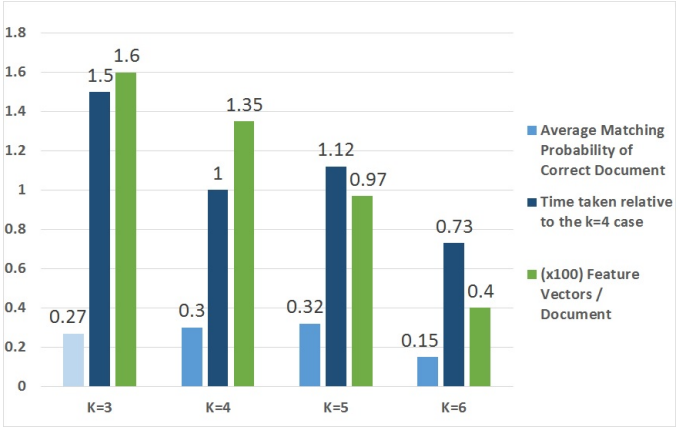


Fig. 5. Experiments with hash function-based retrieval, for the second feature with the probability equation similar to Eq. 2. This figure shows a justification for our using $k = 4$ segments, since it gives a relatively good average matching probability for the correct document, in reasonable time (less than a second on a 2GHz dual core, 2GB memory laptop) showing the average matching probability of the correct document, and the average number of feature vectors per document. Sec. 3 has the details.

document D_i's block d_{ik_j}. We have a set of features $\{f_l\}$ (for our experiments, we have taken two features). We generalise the equation mentioned in Sec. 2.1 (Eq. 1) to that of any feature f_l, since we have a common geometric hashing-based strategy for a feature vector (as in Sec. 2 and Sec. 3). For an individual feature f_l, we compute the probability of an image block q_j actually being database document image block d_{ik_j}:

$$P_{f_l}(d_{ik_j}|\, q_j) = \frac{\text{votes cast for } d_{ik_j} \text{ by } q_j}{\text{total votes cast by } q_j} \tag{4}$$

As mentioned before, we compute the probability values for those blocks which correspond to the same index in the hash table (the collision chain). With independent features f_l (independent sources of measurement), the probability of the observed block q_j being actually d_{ik_j}, given information from f_l, is:

$$P(d_{ik_j}|\, q_j) = \prod_l P_{f_l}(d_{ik_j}|\, q_j) \tag{5}$$

Here, we consider the intersection of the collision chains for the independent hash tables of the features. As such, the information fusion method is independent of the characteristics of specific features f_l used.

The query image has n blocks $q_1, q_2 \ldots q_n$. These could correspond to the database document blocks d_{ik_j}, $j = 1 \ldots n$. Since the evidence from difference blocks are independent of each other, the joint probability of the text blocks d_{ik_j} is:

$$P(d_{ik_1}, d_{ik_2}, \ldots, d_{ik_n}|\, q_1, q_2, \ldots, q_n) = \prod_n P(d_{ik_j}|\, q_j) \tag{6}$$

We can now use the above expression to compute the *a posteriori* probability of document D_i given that we have observed query image Q:

$$P(D_i|\, Q) \propto \sum_{t_i} P(d_{ik_{t_{i_1}}}, d_{ik_{t_{i_2}}} \ldots d_{ik_{t_{i_n}}}|\, q_1, q_2 \ldots q_n) \tag{7}$$

where the summation is over all possible block configuration hypotheses t_i corresponding to document D_i. The given query image Q corresponds to exactly one document D_i, a closed-world assumption in our system. Hence,

$$\sum_{i=0}^{m} P(D_i|\, Q) = 1 \tag{8}$$

Then, using Eq. 7 and Eq. 8, the *a posteriori* probability for document D_i is given by Eq. 9, below

$$P(D_i|\, Q) = \frac{\sum_{t_i} P(d_{ik_{t_{i_1}}}, d_{ik_{t_{i_2}}} \ldots d_{ik_{t_{i_n}}}|\, q_1, q_2 \ldots q_n)}{\sum_z \sum_{t_z} P(d_{zk_{t_{z_1}}}, d_{zk_{t_{z_2}}} \ldots d_{zk_{t_{z_n}}}|\, q_1, q_2 \ldots q_n)} \tag{9}$$

The summation in the numerator is over all hypotheses t_i corresponding to document D_i (as in Eq. 7 above), and that in the denominator is over all hypotheses t_z calculated for all documents D_z.

5. Experiments with Wide Variations in Query Images: Geometric Deformations, Illumination Variations, Noise

In this section, we describe some of our pre-processing modules, and results of the system in handling difficult query images. Our database has 10,000 document images and an equal number of query images. These 10,000 query images correspond to challenging imaging conditions, some with skew alone (3100), some with occlusions (2400), some with imperfect illumination (2600), and some with a combination of one of more of the above (1900). For simplicity, we consider the document with the highest probability value, as the recognised document. We use the term 'accuracy' in the same sense as the work of the Osaka Prefecture University group (which they also consider to be of prime importance), as what is termed 'precision' in precision-recall/sensitivity-specificity/ROC studies. This is the ratio of the true positives to what the system returns as positives i.e., true positives and false negatives.

5.1. Handling Illumination Variations, Other Pre-processing

To handle effects of illumination variation in the query image, we use a relative gradient image (Wei and Lai (2006)), in place of ordinary image intensities:

$$I(x, y) = \frac{|\nabla F(x, y)|}{\max_{(u,v)\in W(x,y)} |\nabla F(u, v)| + c} \tag{10}$$

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 small numbers. Fig. 6 shows a case of successful recognition of the system in spite of bad illumination (here, there is a considerable skew as well) in the query image. Sec. 5.2 delves deeper into how our information fusion scores over using individual features alone. In Fig. 6 and all other results, the block in question appears 'straightened' since for both features, we are matching query blocks with database document blocks, which have zero skew, and are 'straight'. The projective invariants with the contour extrema correspond to estimating the homography which map the given query block to the corresponding

Fig. 6. The system works in cases of bad illumination (here, there is considerable skew as well): an example. The figure shows (left-to-right), the input image; the contour extrema features on the prominent text block; and the zig-zag (word bounding box) features. In both cases (as with all other results), the query block is relatively 'straightened' as the features map the given possibly skewed query image block, to a database document image block, which is in a canonical skew-less orientation. Details in Sec. 5.1.

document block. Similar is the case with the matching angles for the second feature. Other pre-processing modules include binarisation (Lins and da Silva (2007)), smearing (Cao et al. (2007)) for text blobs, and computing curvature extrema points for the first feature (Sec. 2).

5.2. Information Fusion for Robust Matching; Success in Handling Other Difficult Cases

Fig. 7 illustrates the relative merit of using our probabilistic combination of features (Sec. 4), as opposed to using a single feature alone. The accuracy figures for the contour extrema features alone are 73.57% and for the zig-zag (word bounding box) feature is 74.76%, whereas the combination of the two gives an accuracy of 91.49%. Further, Fig. 8 shows the distribution of the relative numbers of query images, corresponding to the maximum probability matching document (for the correct cases alone), for a fraction of the total number of queries. The figure shows that the distribution for the probabilistic combination of features clearly out-performs that for any individual feature.

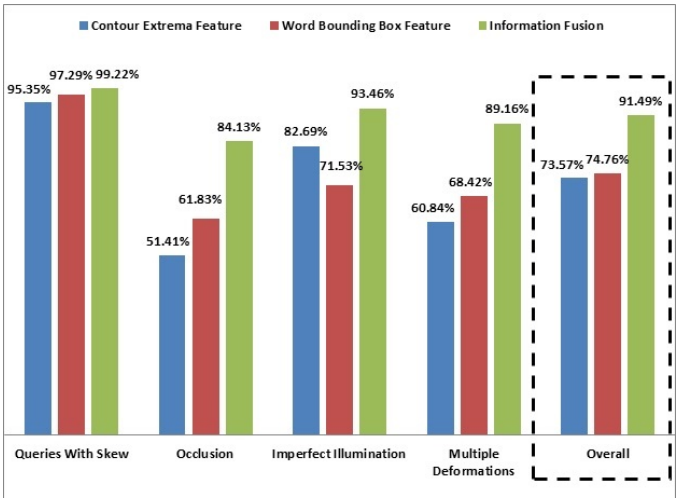


Fig. 7. Performance of probabilistic fusion of information from both features, with special reference to difficult cases

Fig. 9 shows an example of successful matching in spite of a large amount of skew in the query image. For cases of skew as the main deformation, the performance of the contour extrema features alone is 95.95%, the zig-zag (word bounding box) feature is 97.29%, while the probabilistic combination of the two gives 99.22% (Fig. 7). For cases of occlusion alone, the accuracy with the contour extrema features alone is 51.41%. For the

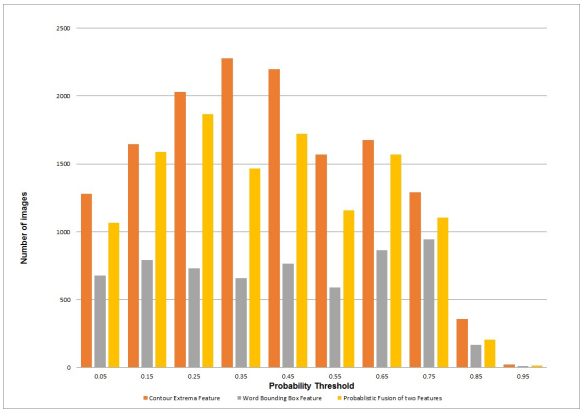


Fig. 8. The distribution of the number of query images corresponding to the probability of the maximum matching document for correct cases alone (bin size=0.10), for each individual feature, and for our probabilistic combination of the two features. The latter clearly gives better results.

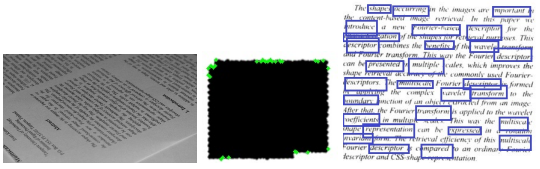


Fig. 9. Figure showing the performance of the proposed technique on a highly skewed query image. From left-to-right, the original query image; its binarised version; and the contour with its prominent curvature extrema points

zig-zag (word bounding box) feature, this number is 61.83%. However, the corresponding figure for the combination of the two features is 84.13%. This clearly shows the utility of the two features taken together, for cases of occlusion. Local information in the form of both boundary information of a block, as well as the structure of words inside a block, play an important role in identifying the particular block, when a large part of either the contour, or the interior of the block, are occluded by external objects. For instance, a metallic object could have been used to keep paper from flying away (as in Fig. 1), or a finger, to hold the document towards the mobile phone camera, as in Fig. 3. In Fig. 10, occlusion from the two metallic objects affects one feature, in this case the zig-zag (word bounding box) feature. Our probabilistic information fusion strategy enables correct recognition. Fig. 11 shows an example with two prominent deformations namely, occlusion, and skew as well. In this case, the occluding object (the key) affects *both* the contour extrema feature, as well as the zig-zag (word bounding box) fea-

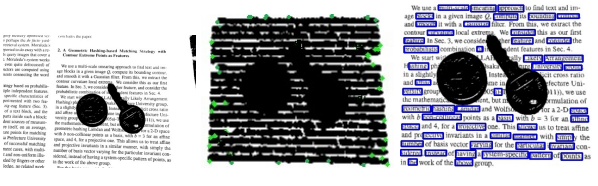


Fig. 10. Successful matching in spite of structured noise/occlusion. In this case, the zig-zag word bounding box) feature is affected by the occlusion, but the feature (contour extrema) is relatively unaffected by it. Our probabilistic information fusion strategy enables successful recognition.

Fig. 11. Successful matching in spite of structured noise/occlusion, and some skew as well. Unlike Fig. 10 where the zig-zag (word bounding box) feature is affected by occlusion, in this case, *both* the contour extrema feature and the zig-zag feature are affected. Our probabilistic information fusion strategy enables successful recognition, again.

Fig. 12. As such, the system does not use any script-specific information. Here, we show results of successful matching on a query image with a mixture of Devanagari and Roman scripts. The first image shows the pre-processed binarised query image, and the rest have the two features, the contour extrema points, and the zig-zag (word bounding box) features, superimposed on the pre-processed image

ture. In this case again, our probabilistic information fusion strategy enables correct recognition.

As such, the system does not use any script-specific information. Fig. 12 shows the results of successful recognition on a document with mixed Devanagari and Roman script. For the 1900 most challenging cases in our small database (multiple deformations in the same query image), the contour extrema feature alone gives a 60.84% accuracy, the zig-zag (word bounding box) alone gives 68.42%, whereas a probabilistic combination of the two gives an accuracy of 89.16% (Fig. 7).

Fig. 13 shows experimental results with a full page document with multiple blocks, and the use of our two features. The probability computations are performed block-wise, with one term for each of the 4 query blocks in Fig. 13, as in Eq. 4, and the probabilistic information fusion is performed as in Sec. 4. For the zig-zag features, we consider a covering of all reasonable-sized word bounding boxes, without overlap (to avoid the combinatorial complexity of considering all permutations).

5.3. Failure Cases

The failure cases are primarily those where the information from the contour extrema features is almost completely unreliable, with most prominent curvature extrema coming from the occluding part, as opposed to the actual contour. This completely pulls down the overall probability value, in spite of a good score from the zig-zag feature. For cases of a fair number of points from the non-occluding proper part of the text blocks, the overall probability stays relatively high. This ensures that the catastrophic failures cases are very low, as borne by the overall high probability of success.

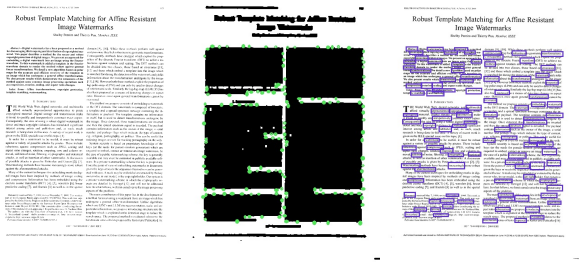


Fig. 13. Results on a full page document with multiple blocks: (from left to right) the pre-processed binarised query image, the contour extrema points, and the and the zig-zag (word bounding box) features

An advantage of the proposed system is the relative independence to the language or script of the document itself. However, some languages may have scripts which are not amenable to the common definition of a word as a sequential combination of characters. This may conflict with the thresholds for the relative size of a 'word'. Fig. 14 shows a case of catastrophic failure for a Chinese language document page. For the sake of illustration, we tweaked the word threshold so that we end up with insufficient character/word spacing. This leads to an almost complete failure of the second feature, which pulls down the overall probabilities below the threshold.

Fig. 14. A case of failure, when the near-failure of a particular feature pulls down the overall probability. For the sake of illustration, we tweaked the bounding box size threshold. For this Chinese language document, the word bounding box feature does not work properly, owing to insufficient word-character separation. (Sec. 5.3)

6. Discussion: A Comparison of the Proposed System, with the State-of-the-Art

As mentioned in Sec. 1, perhaps the state-of-the-art in camera-based document image retrieval is the work of the Osaka Prefecture University group, and the systems they constructed and demonstrated. Sec. 2 points out some important differences between the point features used by this group, and ours (just prior to Sec. 2.1). The retrieval speeds in the Osaka Prefecture University systems range from 1/7 second (0.14ms) on a database of 10,000 pages (Nakai et al. (2006)), to a the memory-reduced and code optimised version of the system (Takeda et al. (2011)), where for a database of 20 million pages, it takes an average of 49ms. In our experiments with a downloaded version of the first system (Nakai et al. (2006)), retrieved in late 2011 from `http://www.m.cs.osakafu-u.ac.jp/~kise/LLAH/index.html` (which uses affine invariants), we have also implemented a non-optimised (for speed and memory) version of the algorithms of the group, extending the ideas to use projective invariants, for a fair comparison. For our database, the downloaded code gave similar real-time retrieval performance, in a small fraction of a second. Our method does not compare well in terms of average retrieval time (being a plain vanilla implementation of the ideas, and not optimised in any way), taking tens of seconds on a 2GHz dual core, 2GB memory laptop.

Our system scores over the Osaka Prefecture University systems in terms of the overall robustness for difficult cases, as

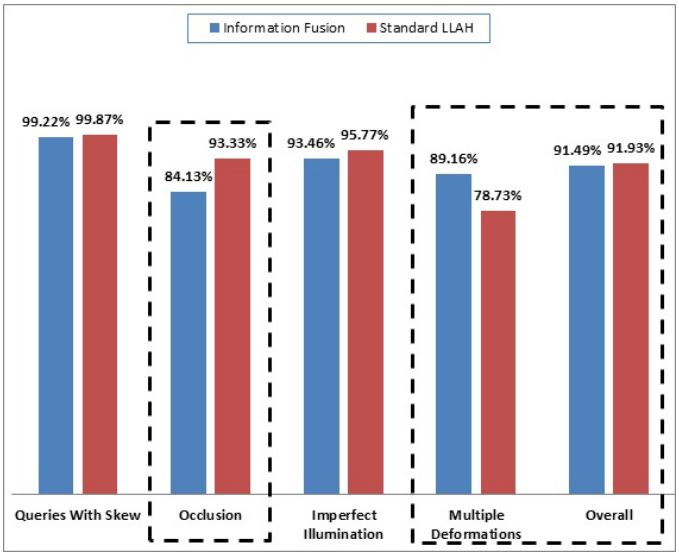


Fig. 15. While our system ('Information Fusion') does not do as well as the Osaka Prefecture University systems ('standard LLAH') in terms of retrieval speed (Sec. 6), our performance is better for difficult cases, such as those shown in Fig. 1). On our representative database, our system performs relatively better on difficult cases, especially on those with multiple simultaneous deformations, such as low resolution, occlusion and imperfect illumination, all in the same image.

shown in Fig. 15. On our representative database, our system ('Information Fusion' in Fig. 15) performs better than the Osaka Prefecture University Systems ('Standard LLAH' in Fig. 15) especially in cases of multiple simultaneous deformations, such as problems related to low resolution, occlusion and imperfect illumination, all in the same image. The overall accuracy performance of the system is quite comparable, but our information fusion system outperforms the Osaka Prefecture University systems with multiple deformation in the image (89.16% in our case, to 78.73% for theirs), as in Fig. 15. As mentioned before, we define 'accuracy' as in the work of the Osaka Prefecture University group (they consider 'accuracy' as being of prime importance in their perhaps *de facto* state-of-the-art document retrieval system), as the relative ratio of the true positives to what the system returns as the result. This would be the 'precision' in precision-recall/sensitivity-specificity/ROC analysis.

Fig. 16 shows two examples of failure of the Osaka Prefecture University systems. These are two representative cases, of an occluding hand over a document (the figure to the left) and some powder sprinkled on a text block, with insufficient illumination as well (the figure on the right). In both cases, the systems fail due to an insufficient number of word centroid points being extracted from the image, corresponding to the correct data from the text blocks. Our experiments also correlate with the observations of Moraleda (2012), which point to failures in cases where a fair part of the query document is not visible in the image. As such, while the system of Moraleda (2012) proposes a scalable system that works well for defocussed images as well, it is not based on full projective invariance, and hence, cannot handle cases of severe document skew.

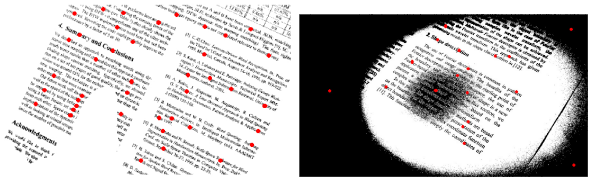


Fig. 16. Some failure cases of the Osaka Prefecture University Systems, showing the feature detection on a pre-processed binarised image, (a) occlusions (from holding a large part of the document with a white glove) resulting in a small part of the actual image being visible, and (b) images with poor resolution, and extremely bad illumination and occlusion (from some black powder sprinkled on the paper). In both cases, the system fails since the very few word centroid points corresponding to the actual data, are extracted from the image.

7. Conclusions

The paper presents a robust document matching system for document retrieval, given challenging query images (insufficient and non-uniform illumination, skew, and occlusions). The system probabilistically fuses information from multiple sources of measurement, which causes the system to work in cases when a particular feature does not give acceptable results all by itself. Experiments using a representative database show encouraging results.

References

Cao, H., Prasad, R., Natarajan, P., MacRostie, E., 2007. Robust page segmentation based on smearing and error correction unifying top-down and bottom-up approaches, in: Proc. ICDAR, pp. 392–396.

Hermann, P., Schlageter, G., 1993. Retrieval of Document images using layout knowledge, in: Proc. ICDAR, pp. 537 – 540.

Hull, J.J., 1994. Document Image Matching and Retrieval with Multiple Distortion-Invariant Descriptors, in: Proc. DAS, pp. 383–400.

Lamdan, Y., Wolfson, H.J., 1988. Geometric hashing: A general and efficient model-based recognition scheme, in: ICCV, pp. 238 – 249.

Liang, J., Doermann, D., Li, H., 2005. Camera-based analysis of text and documents: a survey. IJDAR 7, 84–104.

Lins, R.D., da Silva, J.M.M., 2007. A Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents, in: ASAC, pp. 610–616.

Liu, H., Feng, S., Zha, H., Liu, X., 2005. Document image retrieval based on density distribution feature and key block feature, in: Proc. ICDAR, pp. 1040–1044.

Liu, X., Doermann., D., 2007. Mobile retriever-Finding the Document with a Snapshot, in: Proc. CBDAR, pp. 29–34.

Moraleda, J., 2012. Large scalability in document image matching using text retrieval. PRL 33, 863 – 871.

Nakai, T., Kise, K., Iwamura, M., 2005a. Camera-based document image retrieval as voting for partial signatures of projective invariants, in: Proc. ICDAR, pp. 379 – 383.

Nakai, T., Kise, K., Iwamura, M., 2005b. Hashing with local combinations of feature points and its application to camera-based document image retrieval, in: Proc. CBDAR, pp. 87 – 94.

Nakai, T., Kise, K., Iwamura, M., 2006. Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval, in: Proc. DAS, pp. 541 – 552.

Takeda, K., Kise, K., Iwamura, M., 2011. Memory reduction for real-time document image retrieval with a 20 million pages database, in: Proc. CBDAR, pp. 59 – 64.

Wei, S., Lai, S., 2006. Robust and Efficient Image Alignment Based on Relative Gradient Matching. IEEE Trans. on IP 15, 2936–2943.