Detection of Cut-And-Paste in Document Images

Ankit Gandhi and C. V. Jawahar
Center for Visual Information Technology, IIIT-Hyderabad, India
Email: ankit.gandhiug08@students.iiit.ac.in, jawahar@iiit.ac.in

***Abstract*—Many documents are created by Cut-And-Paste (CAP) of existing documents. In this paper, we proposed a novel technique to detect CAP in document images. This can help in detecting unethical CAP in document image collections. Our solution is recognition free, and scalable to large collection of documents. Our formulation is also independent of the imaging process (camera based or scanner based) and does not use any language specific information for matching across documents. We model the solution as finding a mixture of homographies, and design a linear programming (LP) based solution to compute the same. Our method is presently limited by the fact that we do not support detection of CAP in documents formed by editing of the textual content.**

Our experiments demonstrate that without loss of generality (i.e. without assuming the number of source documents), we can correctly detect and match the CAP content in a questioned document image by simultaneously comparing with large number of images in the database. We achieve the CAP detection accuracy of as high as 90%, even when the spatial extent of the CAP content in a document image is as small as 15% of the entire image area.

Keywords*—document retrieval; linear programming and optimization; plagiarism detection; camera-based document image processing

I. Introduction

With the emergence of large document repositories, many new documents get created by cut and paste (CAP) of documents. Often these could also lead to unethical CAP. Different methods have been employed in the literature to detect such cases, with the focus on plagiarism detection. We believe there can be two directions to detect such cases of document forgeries – recognition based and recognition free. Text level comparison of documents is fairly advanced and there exists many software tools for this [1]. They rely on the OCRs (for text creation) and some level of language processing in the text domain to find similarity between documents. However, text is not always available. In this paper, we focus on a recognition free solution to the problem of detecting CAP in document images. To the best of our knowledge, this is the first attempt on detection of recognition free CAP and plagiarism.

Since our method is recognition free, our closest work is that of retrieving similar documents given a query document, which is popularly known as recognition free document retrieval. Recent years have seen significant progress in document retrieval credited mostly to the development of better representation of document images (e.g. configuration of interest points [2], BoW [3], profile features [4]) and better indexing schemes (e.g. LLAH [2], LSH [5], inverted-index [3]). There are two directions of work in this category. The first category of attempts consider the query as a complete (or at least a large part of) document [2], and the second category focus on query by specific word examples [5]. Because of the advancement of the portable devices and commodity hardwares, research in this direction also needed to support camera based queries [6], [7], as also is Google-Goggles [8].

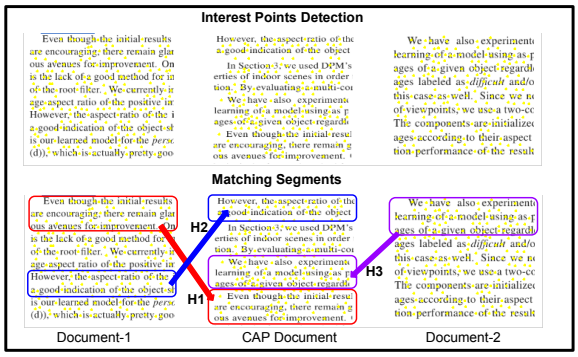

Fig. 1. Detection of Cut-And-Paste in a questioned document image: We are interested in matching parts of CAP questioned document with the documents in database using mixture of homographies model. In the figure, CAP document is matched with Document-1 & Document-2 using 3 homographies.

Our problem, in a way, is a retrieval problem where the query is a part of a document (say a paragraph or few lines) and therefore, falls in between the two categories of work described above, which retrieves results based on query as entire document or query as words. However, there is a significance difference for our problem. Our query may be the entire document but we are interested in retrieving multiple documents from which parts are possibly cut and paste. Our "queries" are part of the documents, but we do not know which parts need to be used as queries. Therefore, our problem can be better modelled as matching parts of the documents to multiple documents present in the database. Since, we do not know the question regions, our method needs to be segmentation free. By modelling the matching as reliably fitting a *homography*, we simultaneously support the camera based (perspective) imaging and traditional scanner based (orthographic) imaging. Solution to fitting one homography between two images is well known [9]. Simultaneously fitting multiple homographies to match parts in different documents is also a novel contribution of this paper. Figure 1 shows the matching of parts of CAP document with parts of Document-1 and Document-2 using multiple homographies.

Document forensics and security has emerged as a prominent research area with immediate applications. A major direction of research in this domain is to validate the authenticity of the handwritten documents. Some of the recent works in the field of forensic science includes signature verification [10], [11], handwriting analysis [12], fingerprint analysis [13], etc. The work has also been done on automatic detection of forged documents by detecting the geometric distortions introduced during the forgery process [14]. In this work, we detect the forgery and plagiarism in documents by detecting CAP content

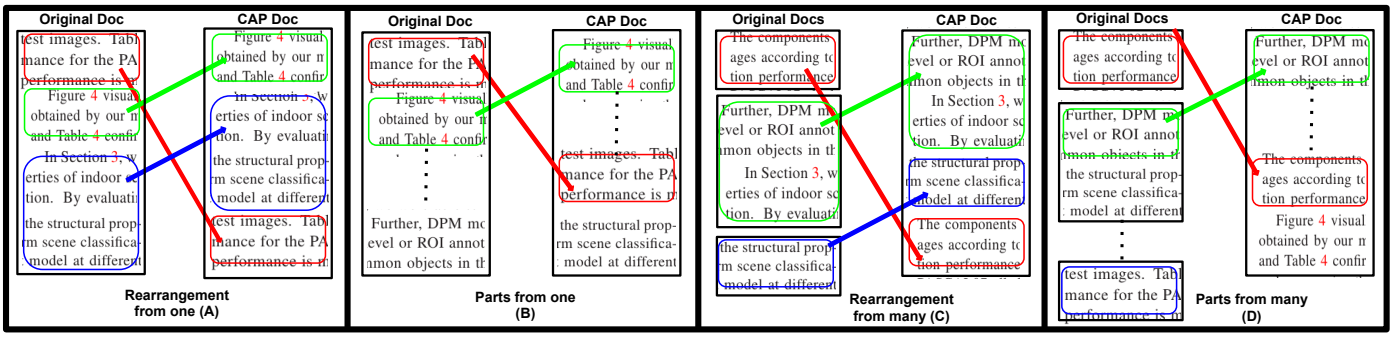

Fig. 2. Different scenarios explaining the creation of CAP documents. (From left to right) (i) CAP document is same as source document with contents being rotated, (ii) Only some part of the source document is copied, (iii) CAP document is the rearrangement of many multiple source documents, (iv) CAP document is created by copying some parts from multiple documents.

in it.

Because of the increasing availability of low-priced digital cameras, many applications have been built over the recent years to work on images captured using camera. The images captured from cameras are of often low quality and suffers from perspective distortion. In most applications, perspective distortion is removed with the help of homography [6], [7] or fundamental matrix [15] so that the same scanner-based document analysis techniques can be applied to camera images also. We also use homography to remove perspective distortion while detecting CAP.

We validate our method by performing experiments on synthetically generated CAP documents as well as on real CAP document. Experiments have been performed by changing the imaging process, varying the percentage of CAP in document image and changing the number of sources from which content has been copied. In all the above scenarios, we are successfully able to detect CAP with a considerable high accuracy. Our method presently does not detect the cases of plagiarism when the text is formatted as this makes the use of geometry very hard.

II. CAP Detection as Fitting Mixture of Homographies

In this section, we discuss the different scenarios of cut and paste in document images. We divided them into four categories depending upon the number of documents from where content has been copied and how much has been copied. They are pictorially explained in Figure 2. (A) *Rearrangement from one* - In this case, CAP document has been created by cutting and pasting another document but its contents have been reshuffled, (B) *Parts from one* - In this scenario, some of the contents of CAP document are original and some of them have been cut and pasted from another document. (C) *Rearrangement from many* - In this case, CAP document has been created by cutting and pasting contents from multiple documents. CAP documents doesn't have any original content in this case. (D) *Parts from many* - This case is a generalization of above case. CAP documents are created by cutting and pasting some parts from multiple document sources although they have some original content also. In all these scenarios, the question/candidate document and database documents may be captured in two different imaging scheme.

Consider the document images in Figure 3. Contents of both the images are same however the images have been

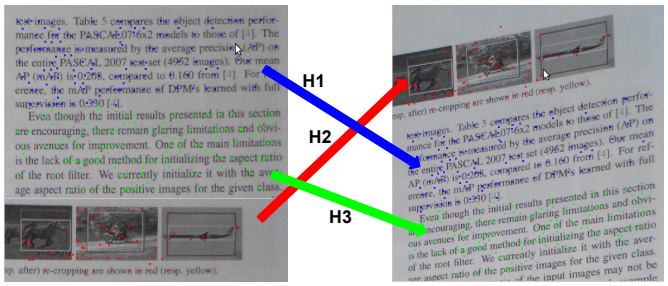

Fig. 3. Invariance to view point variations. Same colour shows points belonging to same homography.

taken at different angles and also geometric transformations have been applied to paragraphs, figures, lines and the other contents in one image. Single homography would suffice if we match two views of the same documents and no CAP have been applied to their contents. In our case, we need multiple homographies to match the two documents (single homography for each part of the image). In this section, we propose a simple LP formulation for spatially matching the two documents using a mixture of homographies model.

The problem is first solved by generating a candidate homographies set and it is assumed that all the true homographies exist in the candidate set. We relax this assumption in the next section. Binary variables are declared for every pair of matching point-pair correspondence between two document images and the homographies in candidate set. Then, the LP is solved to know which homography matches which point-pair correspondence. At last, homographies are re-estimated from the correspondences that have been assigned to them. Unlike the local RANSAC algorithm introduced by Iwamura et al. [16], we use LP to fit the multiple homographies as it seeks to achieve a better optimum value than the former [15].

Given two documents, and a set of interest points (say extracted with SIFT detectors), we can match them and fit a homography, since the point correspondences across these documents will satisfy the relationship $\mathbf{x_i} = H\mathbf{x'_i}$. Given enough matches, one can compute a robust estimate of this homography by minimizing an error function of the form

$$\sum_i d(\mathbf{x_i}, H^{-1}\mathbf{x'_i})^2 + d(\mathbf{x'_i}, H\mathbf{x_i})^2, \qquad (1)$$

where $d(\mathbf{x}, \mathbf{y})$ denotes the Euclidean distance between points represented by $\mathbf{x}$ and $\mathbf{y}$.

However, in our problem this direct solution does not

work. We want the matches across multiple parts of document. This naturally imply that we need to find multiple homographies simultaneously. For this purpose, we introduce a binary variable z_{ik} to denote whether ith correspondence is part of the kth homography or not. Assume we have a list of M candidate homographies, denoted by $\phi = \{H_1, H_2, \ldots, H_M\}$. We assume that all the true homographies are present in the candidate homographies set. Let $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ be the point correspondences. We now define the error function as :

$$\sum_{i=1}^{n}\sum_{k=1}^{M} z_{ik}(d(\mathbf{x_i}, H_k^{-1}\mathbf{x_i'})^2 + d(\mathbf{x_i'}, H_k\mathbf{x_i})^2) = \sum_{i=1}^{n}\sum_{k=1}^{M} z_{ik}d'_{ik},$$

where $d'_{ik} = (d(\mathbf{x_i}, H_k^{-1}\mathbf{x_i'})^2 + d(\mathbf{x_i'}, H_k\mathbf{x_i})^2)$.
Let y_k $(k = 1, \ldots, M)$ denotes the binary variables indicating whether H_k is one of the true homographies. Similar to the objective function defined in [15], we also have the data term and complexity term in our formulation explaining how well the homography model fits the data and how complex our models are respectively. Minimization of the error function can be modelled as an integer linear program :

$$\min_{\mathbf{z},\mathbf{y}} \sum_{i=1}^{n}\sum_{k=1}^{M} z_{ik}d'_{ik} + \beta\sum_{k=1}^{M} y_k, \tag{2}$$

subject to the constraints:

$$A : \sum_{k=1}^{M} z_{ik} = 1 \quad B : \max_{i=1}^{n}\{z_{ik}\} = y_k \quad C : z_{ik}, y_k \in \{0, 1\}.$$

Constraint A assures that every matching point-pair correspondence can belong to only one homography of the candidate set and constraint B allows that correspondence i can be assigned to homography k only if H_k exists in the true homography set. The above proposed formulation is binary integer linear programming which is hard to solve in practice. LP-relaxation is one of the known and effective methods to solve such problems in which binary variables are replaced with real continuous variables. We have used LP-relaxation to solve the above formulation. This provides a soft assignment solution.

Once the membership of all matching point-pair correspondences is known i.e. the homography to which each correspondence belongs, the homographies are re-estimated with all the correspondences that belongs to it using RANSAC algorithm (Equation 1).

III. Extensions and Refinements

A. Handling outliers

The above formulation has limitation that it assigns every matching point-pair correspondences in documents to a homography. But in practical scenarios, we have lot of false positives in point-pair correspondences, e.g., In document images, because of presence of stop words such as *the, who, is* and repetition of characters there will be lot of false point-pair correspondences. In general, we want our model to ignore all such correspondences and learn homography from true point correspondences. Such false point-pair correspondences are known as outliers and they do not correspond to any homography. In order to make our formulation robust to outliers, we introduced another set of binary variables w_i $(i = 1, \ldots, n)$, for each point-pair correspondence to denote whether it is an outlier or not. The variable w_i=1 indicates ith correspondence is an outlier and w_i=0 otherwise. Following is the formulation of mixture of homographies incorporating outliers -

$$\min_{\mathbf{z},\mathbf{y},\mathbf{w}} \sum_{i=1}^{n}\sum_{k=1}^{M} z_{ik}d'_{ik} + \beta\sum_{k=1}^{M} y_k, \tag{3}$$

subject to the constraints:

$$A : \sum_{k=1}^{M} z_{ik} + w_i = 1 \qquad B : \max_{i=1}^{n}\{z_{ik}\} = y_k$$

$$C : \sum_{i=1}^{n} w_i \leq P \qquad D : z_{ik}, y_k, w_i \in \{0, 1\}$$

Constraint P puts upper bound on the number of outliers in order to avoid trivial solutions. We used the above robust formulation in all our experiments. Note that while matching interest points in document images, number of outliers are very high as explained above, however, our formulation can robustly ignore all such false matchings.

B. Candidate Set Generation

In Section II and III-A, we have made strong and crucial assumption that the set of candidate homographies are known to us beforehand. In this section, we relax that assumption. We propose an efficient technique to generate such a candidate set. Note that four matching point-pair correspondences are enough to estimate a homography. One simple naive method to generate such a candidate set is to take different sets of four random matching point-pair correspondences and estimate homographies from them. Although this technique works well but is computationally very expensive as the size of candidate set will be very large if we want to ensure that it contains all the true homographies that exist in the image.

We use a variant of above method which also takes into account spatial information of the coordinates of matching point-pair while estimating homographies as there is a high probability that the neighbouring correspondences will also belong to the same homography. Once the correspondence between images are known, we partition the images into 4 quadrants. For each of the quadrants in image, depending on where its corresponding matching point-pairs lie in the second image homographies are estimated. If it spans across multiple quadrants in the second image, then all of them are considered one-by-one for estimating homographies. Using this, the search space from where random correspondences are picked for estimating homographies is reduced. Therefore, size of candidate set using this method is very less as compared to the naive one.

C. Scalability

The naive approach for extending the above method to large number of documents would be to apply the proposed LP solution to every document in the corpus. But this will be computationally very expensive and the standard method in literature to avoid this or to prune the candidate list is to use Bag of Words (BoW) model [17]. The BoW model quickly filter out related images in the corpus when implemented efficiently using reverse index table. After this, LP can be solved for the

TABLE I. CAP DETECTION ACCURACY ON A LARGE CORPUS

CAP Setting	A	B	C	D	Mean
Detection Accuracy (%)	94.4	93.6	85.2	83.4	89.2

top-a documents obtained to find CAP content in the questioned document.

Firstly, SIFT descriptors are computed at interest points for each document image in the corpus. A sample of the descriptors are clustered using approximate k-means algorithm to form a vocabulary of visual words. Every descriptor of the images in corpus is now quantized/mapped to the visual word nearest to it and then used to index the images. Inverted file index is created which has an entry for each visual word storing a list of all the document images containing that word. Given the CAP questioned image, all the visual words present in it are obtained. With the help of reverse index table, all the documents containing common visual words with the questioned document are obtained. Every document image is represented using a M (vocabulary size) dimensional vector $< t_1, t_2, \ldots, t_M >$ where each term is weighted using tf-idf -

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i}, \tag{4}$$

where n_{id} is the number of occurrences of visual word i in document image d, n_d is the total number of visual words in document image, N is the total number of document images in corpus and n_i is the number of occurrences of visual words i in the whole corpus. The questioned document image is also represented using M dimensional vector and its similarity with other document images is calculated using cosine similarity and then the images are ranked using calculated score. Top-a are then selected for solving LP. This pruning can result in missing of original document for matching. However, as we validate in next section, this rarely happens.

IV. RESULTS

In this section, we demonstrate the quantitative performance of the proposed approach in multiple settings. Matching point-pair correspondences between two images is found using FLANN library [18]. Firstly, possible matching point-pair correspondences is generated by finding 5 nearest neighbour of all descriptors of the CAP questioned document in the document with which it is matched. Then, the set is trimmed such that it contains only good matches, all the matches whose distance is greater than threshold are ignored. MOSEK[1] library is used for solving linear programs.

A. Scalability

We first show the effectiveness of proposed method to detect CAP in a questioned document by matching it with documents from the large corpus. For the experiments, we have considered a dataset of approximate 10,000 scanned images, taken from multiple English novels, each of size 2088 $\times$ 3524. A CAP questioned image is generated by cut and pasting random content/lines from multiple documents in the database. We generate 100 such CAP documents for each of the settings discussed in Section II. Given a questioned document, we are interested in obtaining all the documents from the database from where contents have been copied and spatially mapping the CAP content in each of them.

[1]http://www.mosek.com/

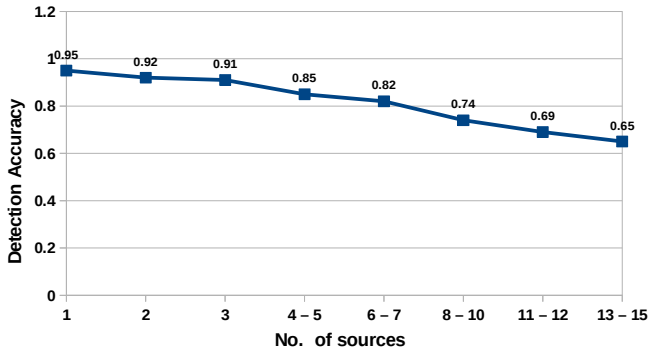

Fig. 4. Variation of detection accuracy with number of sources from which CAP document has been created.

Using BOW retrieval scheme as described in Section III-C, top-100 document images are obtained from the database for a given CAP questioned document. Vocabulary size is taken as $M = 20,000$. Now, LP is solved for each of the probable (top-100) document images and using mixture of homographies model, CAP contents are spatially mapped between a questioned document and the documents from corpus. P (upper bound on the number of outliers) in (3) is fixed as 25% of the n (total number of matching point-pair correspondences).

We have measured the performance of our method for each of the settings A, B, C, D discussed in Section II. We evaluate our method using a measure typically used for object detection task – overlap ratio of ground truth and predicted CAP. It is equal to the ratio of two areas – intersection and union of ground truth and predicted CAP. If this overlap ratio exceeds threshold (0.5), we say that the CAP has been identified and detected correctly. Table I shows the average detection accuracy on 100 CAP questioned document images over all the different settings. It must be noted that all the detection accuracies have been obtained at a fixed false positive rate of 0.005%. On an average, we are able to detect 89.2% of the CAP in questioned documents.

B. Variation with number of sources and view point

In this section, we have analysed the detection accuracy when the number of sources from which the content has been copied varies in a CAP questioned image. When the number of source document increases, the number of homographies using which a CAP questioned document has to be matched also increases. Figure 4 shows the variation of detection accuracy with the number of sources from which content has been copied. We are able to achieve the accuracy of as high as 74% even if the CAP questioned document has been copied from 10 different sources which further demonstrates the effectiveness of the proposed approach. The proposed method for detecting CAP is robust enough to handle the view point variations between CAP questioned document and the document with which it is matched. Using of SIFT-BOW pruning and mixture of homographies model makes our approach invariant to view point changes. Figure 3 shows an example of document matching and the corresponding homographies obtained when the view point is changed.

C. Variation with size of CAP content

We now analyse the variation of accuracy of our method with the amount/size of CAP content. For example, if only one

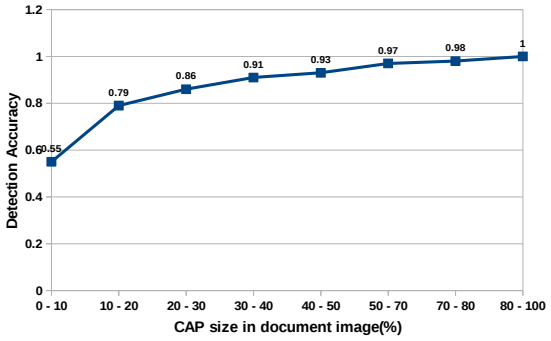

Fig. 5. Variation of detection accuracy with size of CAP content in questioned document.

line or word is copied from a document, then it is hard to detect such scenarios and if one paragraph or more has been copied, then it is easy to detect. Figure 5 shows the accuracy measure with variation of size of CAP content. As expected the problem is relatively hard when size of CAP content is very small. However, when the size of CAP content is 15% or more than that we are able to correctly detect such scenarios with an accuracy of 91.2%.

D. Discussions

In this section, we show a practical application of the proposed method for detecting CAP in two highly similar research papers[2]. We added the first paper[3] to our database of 10,000 documents and a part from the second paper[4] is given as a query document image. We run our CAP detection algorithm for the query document image and our method is able to correctly detect the CAP in it. Figure 6 shows the CAP content matched between the query segment and the document from database using our proposed method. Although the formats of the two paper are different but still our method is able to detect CAP. Also, using our method, we can find percentage of the cut and pasted content in the journal and the conference version of the same paper.

A drawback of the present work is its two step nature to solve the problem in large data sets. In the first step, we use an indexing scheme for building a set of candidates and in the second step, we use geometry to finalize the matching. In future, we wish to do both in the same step, with the help of appropriate indexing schemes. One of the main challenges we see at this stage is to work with documents when the text is reformatted and edited (for example, the text is formatted in a text editor with different line breaks, text widths and font/size variation.). This makes the use of geometry very hard. We are working on this problem in a more flexible Bag of Words setting.

V. Conclusion

In summary, we proposed a LP formulation to detect Cut-And-Paste in document images. Filtering of documents in a large database is done using BoW retrieval approach and then, the CAP content is identified by matching the top documents

[2]http://academicsfreedom.blogspot.in/2012/07/plagiarized.html

[3]http://www.computer.org/csdl/trans/tm/2012/01/ttm2012010086-abs.html

[4]http://www.ijmra.us/project%20doc/IJMIE_JULY2012/IJMRA-MIE1412.pdf

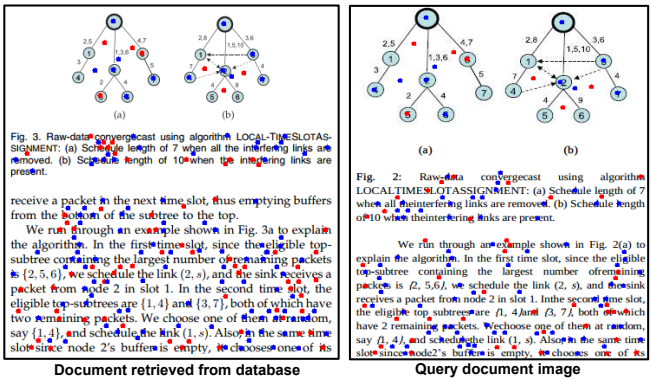

Fig. 6. CAP detection in two highly similar research papers. Papers are matched using two homographies. Figure only shows the snippet of the paper from the database, which gets matched with the query.

with the CAP questioned document using mixture of homographies model. Our experiments confirm the effectiveness of the proposed approach, we are able to achieve considerable high accuracy when number of sources from which CAP document has been created is large and even when the percentage of CAP content in a document is less. The proposed formulation is also robust to handle the outliers and view point variations.

References

[1] [Online]. Available: http://textmatching.com/, http://comparesuite.com/

[2] K. Takeda, K. Kise, and M. Iwamura, "Real-time document image retrieval for a 10 million pages database with a memory efficient and stability improved llah," in *ICDAR*, 2011.

[3] R. Shekhar and C. V. Jawahar, "Word image retrieval using bag of visual words," in *DAS*, 2012.

[4] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *IJDAR*, 2007.

[5] A. Kumar, C. V. Jawahar, and R. Manmatha, "Efficient search in document image collections," in *ACCV*, 2007.

[6] K. Takeda, K. Kise, and M. Iwamura, "Real-time document image retrieval on a smartphone," in *DAS*, 2012.

[7] T. Nakai, K. Kise, and M. Iwamura, "Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval," in *DAS*, 2006.

[8] [Online]. Available: http://www.google.co.in/mobile/goggles/

[9] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[10] V. Blankers, C. Heuvel, K. Franke, and L. Vuurpijl, "ICDAR 2009 signature verification competition," in *ICDAR*, 2009.

[11] M. Malik, S. Ahmed, A. Dengel, and M. Liwicki, "A signature verification framework for digital pen applications," in *DAS*, 2012.

[12] S. Srihari, "Evaluating the rarity of handwriting formations," in *ICDAR*, 2011.

[13] C. Su and S. N. Srihari, "Evaluation of rarity of fingerprints in forensics," in *NIPS*, 2010.

[14] J. van Beusekom and F. Shafait, "Distortion measurement for automatic document verification," in *ICDAR*, 2011.

[15] H. Li, "Two-view motion segmentation from linear programming relaxation," in *CVPR*, 2007.

[16] M. Iwamura, T. Kobayashi, and K. Kise, "Recognition of multiple characters in a scene image using arrangement of local features," in *ICDAR*, 2011.

[17] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.

[18] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *VISSAPP*, 2009.