Key-region Detection for Document Images – Application to Administrative Document Retrieval

Hongxing Gao*, Marçal Rusiñol*, Dimosthenis Karatzas*, Josep Lladós*,
Tomokazu Sato†, Masakazu Iwamura† and Koichi Kise†
*Computer Vision Center, Dept. Ciències de la Computació
Edifici O, Univ. Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain.
†Dept. of CSIS, Graduate School of Engineering
Osaka Prefecture University
1-1 Gakuen-cho, Naka, Sakai, Osaka, 599-8531 Japan.

***Abstract*—In this paper we argue that a key-region detector designed to take into account the special characteristics of document images can result in the detection of less and more meaningful key-regions. We propose a fast key-region detector able to capture aspects of the structural information of the document, and demonstrate its efficiency by comparing against standard detectors in an administrative document retrieval scenario. We show that using the proposed detector results to a smaller number of detected key-regions and higher performance without any drop in speed compared to standard state of the art detectors.**

I. Introduction

Over the past decades computer vision has experienced a rapid growth, one of the main reasons being the vast collections of natural images that were made gradually available. Along with the growth of imagery data and the impracticality of associating meta-data to the images, the need for image based indexing and retrieval has resulted to a plethora of recent advancements with the emphasis being on reducing the size of image descriptors without compromising retrieval efficiency [1].

Key-point correspondence based algorithms have also been used for image retrieval, either through a bag-of-words framework to extract global image descriptors, or by direct key-point indexing in cases when part-based matching is significant. Such approaches are based on a variety of key-point and key-region detectors (e.g. Harris corner [2], Harris-Laplace and Hessian-Laplace [3], Difference of Gaussians [4], Hessian determinant [5], MSER [6], etc.) and an even larger number of local descriptors (e.g. SIFT [7], GLOH [8], SURF [9], HoG [10], etc.).

Although in the document analysis domain there is a chronic lack of large public datasets, the issue of retrieval in big collections of documents has always been a topic of interest with clear socio-economic impact especially in the administrative and the historical document analysis areas. Following suite from the domain of natural images, state-of-the-art key-point detectors and local descriptors have been successfully used in document analysis for document representation in classification and retrieval scenarios [11], as well as other applications such as logo spotting [12], etc.

The basic premise of key-point detectors such as SIFT and SURF is to detect as many stable key-points as possible in order to "densely cover the image over the full range of scales and locations" [13]. Although this makes a lot of sense for object recognition in individual cluttered scene images, it is not necessarily optimal for retrieval applications. Indexing large numbers of local features extracted from an equally large number of images is inefficient, even though it can become tractable through the learning of small codebooks [14] and state of the art hashing and searching techniques [15][16]. At the same time, document images are distinctly different to natural scenes as documents have an explicit structure and are generally high contrast images (giving rise to numerous stable key-points). Classically detected key-points, although they work reasonably well since the densely cover the document image, do not carry any particular semantic or structural meaning.

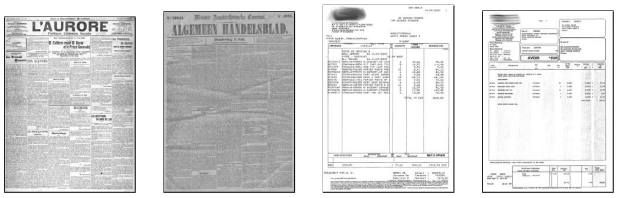


Fig. 1. Typical applications of document retrieval include historical and administrative document analysis uses. (a) images taken from the IMPACT historical newspapers dataset, where a typical application is the retrieval of front pages - used with permission, (b) images from a typical digital mailroom page flow, a typical application being the retrieval of invoices from the same provider.

On the other hand, methodologies specifically designed for document images, make explicit use of document characteristics in their representations. As an example the document matching approach of Nakai et al. [17], makes use of structural features of the document and local topological information. In the case of [17] the centres of blobs detected through blurring and subsequent thresholding -assumed to correspond to words- comprise the key-points, while an affine invariant descriptor encoding the relative position of such blobs in their neighbourhood of the key-point is subsequently used. The indexing and retrieval scheme employed is extremely fast, able to retrieve at 40ms in a dataset of 10 million pages [18].

1520-5363/13 $26.00 © 2013 IEEE
DOI 10.1109/ICDAR.2013.53

The approach of Nakai et al. [17] is indeed a very efficient solution given that the objective is exact document matching. More often than not though, what is of interest is the retrieval of similar documents, and not exact matches. Similar documents might share whole paragraphs of text -in which case word blobs and a feature based on the relative positioning of words could provide a good basis for similarity search- but frequently, similarity is evident in the document structure but not in the exact content. See for example the documents in Figure 1. In our case, we want to retrieve the invoices that are generated by the same provider which might not be similar in terms of their textual content, but still look visually similar.

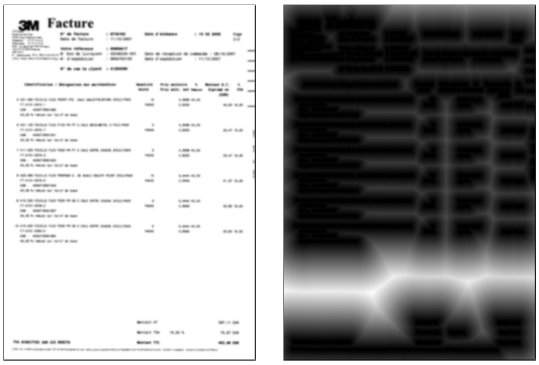
A solution to the problem could be provided by document layout based descriptors, but in reality such approaches are impractical. This is both because of the inherent difficulties of layout analysis (still an open problem, difficult to achieve a repeatable, generic layout analysis algorithm) and because a layout based descriptor is prohibitively expensive to calculate for large datasets. Therefore a compromise should be sought, that bridges the gap between an efficient document descriptor and one that encodes certain structural information of the document.

In this work, we present the first steps towards such a document representation. We focus on the efficient detection of meaningful key-regions that encode structural information among different levels (letters, words, paragraphs and so on). We demonstrate that the proposed key-region detector is efficient to calculate and results to a smaller number of more meaningful regions than other state-of-the-art key-point and key-region detectors such as SIFT and MSER. To demonstrate the suitability of the proposed key-region detector for document retrieval we calculate SIFT descriptors over the detected key-regions and use them for indexing and retrieval in an administrative document scenario. We show that retrieval based on key-regions detected with the proposed method yields better results than other state of the art key-point and key-region detectors.

The rest of this document is structured as follows. In Section II, we study the behaviour of state-of-the-art detectors (SIFT and MSER) when applied in the document image analysis domain. In Section III, we propose a new detector called Distance Transform based MSER (DTMSER) aimed at extracting relevant regions in document images. The experiment results are discussed in Section IV and concluding remarks are given in Section V.

II. Key-Region Detectors in Document Analysis

The standard local feature extraction pipeline that consists of a key-point detector followed by a local descriptor has yielded high performance in numerous challenging problems such as object recognition, robotic mapping and navigation, image stitching, 3D modelling, natural scene understanding, etc. Various detectors (e.g. Harris, Hessian, SIFT, MSER) and descriptors (SURF, HOG, SIFT) have been proposed during the past decade. In this paper, we will concentrate on the performance of three different detectors (SIFT, MSER and our proposed DTMSER) when applied in the document image domain. The SIFT descriptor will be invariably employed to extract local features.



a) b)

Fig. 2. Demonstration of the distance transform. a) Original image, b) its distance transform

In the SIFT framework, key-points are defined as maxima and minima of the difference of Gaussians function applied in scale space to a series of smoothed and resampled images. It therefore detects salient and meaningful blobs and their best representative scale. However, when used in (usually binary) document images, the extrema of the DoG function is usually found at the lower scales, provoking that most of the extracted keypoints correspond to character corners, edges and spaces between characters, instead of higher-level entities. Such key-points are very stable, but present relatively low discriminatory power.

Concerning MSER, key-regions are extracted in terms of the stability of the intensity function over their outer boundary. As such, the algorithm detects blobs that present an important intensity change to their immediate surroundings. When used with document images, the set of maximal regions generally correspond to text parts (usually individual characters) and other dark foreground regions and the set of minimal regions to their white background counterparts. In the extreme case of bi-level images, the output of MSER is roughly equivalent to a connected component analysis. In the document analysis domain, MSER regions have been shown to perform well in matching tasks when dealing with “graphical” documents such as manga [19] comics.

III. Distance Transform based MSER

In the domain of document analysis, it is desirable to identify key-regions that relate to the structural elements of the document, namely characters, words, lines and paragraphs, as they carry important semantic information . Moreover, this should be done in an efficient, repeatable and stable way, as opposed to existing layout analysis approaches which are generally exhaustive and inherently unstable.

The notion of scale in the case of documents is tightly linked to the distance between the structural elements of the document. Characters are placed closer to each other than words are, which are in turn placed closer to each other than paragraphs or columns are. Moreover, the hierarchy of these structures is well defined and informative. On the other hand, the MSER algorithm provides an efficient multi-scale analysis framework, based on the stability over a given pixel property, typically its lightness. The key idea of the detector we

propose is to leverage the efficiency of the MSER algorithm to identify stable regions, where stability is defined as a function of the distance of a region to neighbouring ones. Hence in our framework regions that have larger distances to neighbouring ones would be more stable than regions that are close to each other.

The above algorithm is practically equivalent to a graph contraction approach, over a graph that encodes the neighbouring relationships between the connected components of the image, which in the generic case could be the fully connected graph of the connected components. A graph contraction implementation is quite inefficient. Using instead the distance transform we translate the problem from the distance domain to the image domain, where the MSER segmentation offers an efficient way to create and rank (in terms of their stability) the regions corresponding to clusters of neighbouring connected components.

A. Distance Transform

The distance transform finds the minimum distances of all image pixels to the set of foreground pixels. The result is a matrix of the same size as the image, where each element is assigned a value corresponding to the smallest distance between the corresponding image pixel and the closest foreground object.

We compute the distance transform of the document image based on the two pass algorithm proposed in [20]. Formally, let p be a background point and q a point in the set of foreground objects Q. The distance transform $f(p)$ assigns at each background point p its distance to the nearest object point by:

$$f(p) = \min_{q \in Q} d(p, q)$$

where $d(p, q)$ is the Euclidean distance between background point p and object point q. An example of the distance transform matrix of an administrative document is shown in Figure 2.

Note that we implicitly assume in this discussion that the image is bi-level . However, we should point out that the distance transform concept is readily applicable to grey scale images [21].

B. MSER detection

The set of maximal regions produced by the MSER algorithm is the set of all connected components produced over all possible thresholdings of the input image (essentially identical to a watershed algorithm). When calculated over the distance transform result, the maximal regions roughly correspond to semantically important structures of the document (characters, words, text lines, paragraphs), as can be appreciated in Figure 3.

Applying the MSER algorithm to the distance transform image produces a dendrogram of maximal regions. The leaf regions correspond to the foreground objects, while the mergers in the dendrogram depend solely on the distance between the maximal regions. An example of the typical dendrogram produced is shown in 4.

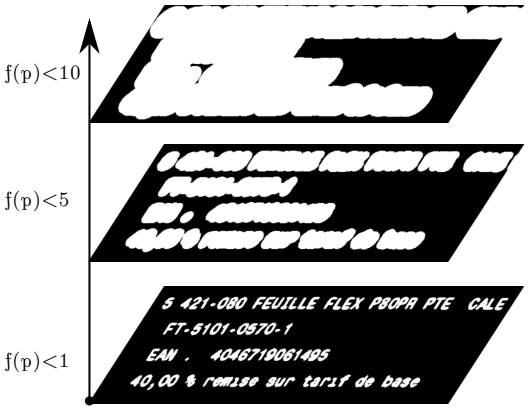


Fig. 3. Example of thresholding the distance transform at different intensity levels. At the lower level individual characters can be identified, at the middle level characters have been merged into words, at the upper level words have been merged into paragraphs.

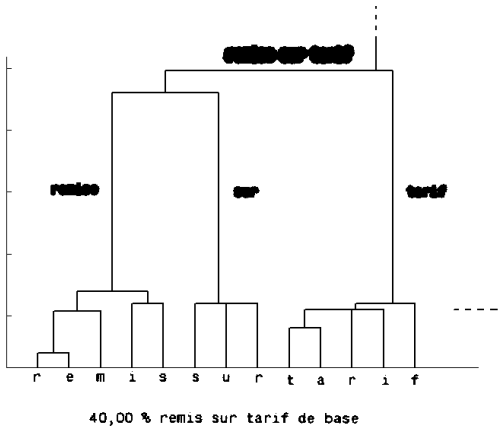


Fig. 4. A typical dendrogram produced with the proposed method. The leaf nodes correspond to the connected components of the image, while the mergers depend solely on the distance between regions, giving rise to semantically relevant groups.

The MSER algorithm's δ parameter controls the minimum lifetime (number of iterations that a maximal region has to survive before merging with a neighbouring one) in order for a region to be considered stable. In the case of documents, and given the prior distance transform, δ effectively controls the minimum distance that a region has to have to a neighbouring one in order to be considered stable. As characters are the most closely positioned structures of interest, we should choose a value for δ that is less than half the minimum distance between characters, in order for them to be identified as stable. In practical terms, we can directly set $\delta = 1$, as we do not expect to have any components positioned closer to each other than characters in the document.

One potential drawback of our proposed detector is the inconsistency of the distance transformation regarding noise. However, coupling the distance transformation with a MSER analysis allows us to extract various key-regions of different sizes, only a small subset of which would be affected by such artifacts.

IV. Experimental Results

We tested our proposed key-region detector in an administrative document retrieval scenario. Our dataset consists of 4109 binary invoice images corresponding to 249 different providers. Using a leave-one-out strategy, given a query invoice we want to retrieve similar documents from our dataset, i.e. invoices from the same provider. Key-regions are detected with SIFT, MSER and the proposed DTMSER methods and are subsequently described by the SIFT descriptor. For each local feature we retrieve the 100-nearest neighbours, each of them casting a vote at document level. For the final document retrieval, documents are sorted according to the votes received and we report the mean average precision mAP and the corresponding precision and recall plots.

We tested three different voting schemes. A *uniform* voting paradigm that gives equal score to all matched documents. An *inverse rank* scoring that weights the document votes depending on their position in the ranked list. Finally, a *truncated inverse distance* scoring function that equally votes for the documents that hold very small distance with the query feature and scores the rest with their inverse distance.

A. Qualitative results

We can see in Figure 5 a qualitative comparison of the types of key-regions identified by the three different detectors. The interest points extracted by SIFT are mainly located at letter corners which may lead to random uncertainty text in the described patch. Most of the MSER produced key-regions correspond to character-level connected components. In contrast, the proposed DTMSER detector extracts multi-level features corresponding to letters, words, and paragraphs which are potentially more semantically meaningful.

B. Comparative results on a subset

We first report comparative results obtained on a subset of the database corresponding to 857 images from 50 different providers in which 10 invoices are selected as queries. To exhaustively find key-region correspondences quickly becomes infeasible when dealing with large datasets. As the SIFT and MSER detectors return an enormous amount of key-regions, we perform this first experiment on a subset of the dataset for computational cost. Furthermore, we make a use of an approximate nearest neighbour search algorithm, namely the Bucket Distance Hashing (BDH) to further reduce the computational time. Bucket Distance Hashing (BDH) is a scalable approximate nearest neighbour search (ANNS) method [22]. The key idea of BDH is a combination of hash-based distance estimation and loose selection of nearest neighbour candidates, both of which are designed to find the true nearest neighbour in high probability without time consuming process. Previous experiments have shown that the BDH can reduce processing time significantly while maintaining the same accuracy as other the state-of-the-art algorithms [23], a behaviour confirmed here as well.

We compare the performance of the three key-region detectors using the three different voting schemes described before. The obtained results are summarized in Table I.

It can be easily appreciated that the amount of obtained key-regions is drastically reduced when using the proposed DTMSER detector instead of SIFT or MSER while the retrieval performance is not affected. There is a clear advantage in using an approximate nearest neighbour search algorithm such as BDH in the retrieval stage as there is a huge time improvement while suffering an insignificant loss in mean average precision. We show in Figure 6 the precision and recall plot for this experiment when using the truncated inverse distance scoring method.

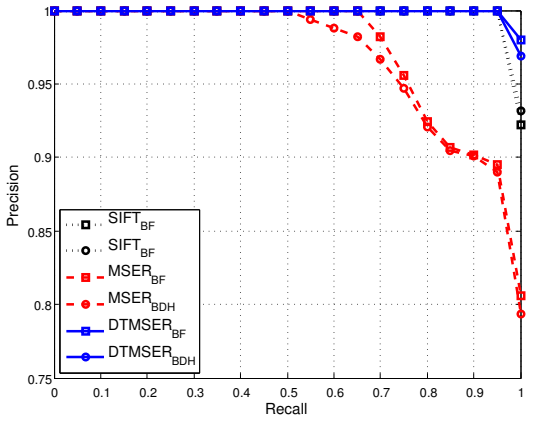


Fig. 6. Precison-Recall curve for the sub-dataset when using the truncated inverse distance scoring method.

C. Results on the whole dataset

To show the performance of the proposed DTMSER, we generalize our experiment over the whole invoice dataset consisting 4109 images within 249 unbalanced classes from which 383 queries are randomly picked. In this scenario the amount of key-regions returned by the SIFT and MSER detectors is very large (SIFT detects more than 40 million key-points). Therefore, we just evaluated the DTMSER detector performance combined with the BDH search algorithm.

We can see in Table II that the performance over the whole dataset is in agreement with what we obtained during the previous experiment. Regarding the different voting schemes, the truncated inverse distance strategy is the one that performs the best, although no significant differences can be observed.

TABLE II. MAP time consumption for whole dataset

Detector	Num. Key-regions	Time (ms)	Uniform	Inverse Rank	Truncated Inverse Distance
DTMSER	2,016,286	205	0.9893	0.9407	0.9909

V. Conclusions

We presented fast and efficient key-region detector that is suitable for document images. The proposed DTMSER detector takes advantage of the particular structure of document images, and is able to detect semantically meaningful key-regions that roughly correspond to structural elements of the document. Compared to other state of the art detectors, DTMSER detects a much smaller number of key-regions, while achieving slightly higher performance in a retrieval scenario.

a) b) c)

Fig. 5. Qualitative comparision of a) SIFT, b) MSER and c) DTMSER key-region detectors.

TABLE I. MAP AND TIME CONSUMPTION FOR SUB DATASET

Detector	Num. Key-regions	NN Search	Time ($m.s$)	Voting Scheme		
				Uniform	Inverse Rank	Truncated Inverse Distance
SIFT	9,402,479	BF	6,148,880	0.9830	0.9955	0.9963
		BDH	3,640	0.9768	0.9945	0.9968
MSER	1,164,693	BF	135,896	0.9654	0.9667	0.9645
		BDH	679	0.9595	0.9658	0.9601
DTMSER	422,288	BF	21,699	1.0000	0.9634	0.9990
		BDH	131	1.0000	0.9659	0.9984

The approach followed produces a dendrogram of regions. In the current implementation all regions produced are indiscriminately used for indexing. We should nevertheless stress that the dendrogram produced is a rich source of structural information, as it encodes relationships between the regions. In future work we plan to examine ways to take advantage of this structural information to further improve the retrieval system.

ACKNOWLEDGEMENTS

This work has been supported by the Spanish projects RYC-2009-05031, TIN2011-24631, TIN2009-14633-C03-03, and China Scholarship Council grant (No.2011674029).

REFERENCES

[1] A. Gordo and F. Perronnin, "Asymmetric distances for binary embeddings," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 729–736.

[2] C. Harris and M. Stephens, "A combined corner and edge detector," in *In Proc. of Fourth Alvey Vision Conference*, 1988, pp. 147–151.

[3] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proceedings of the 7th European Conference on Computer Vision*, 2002, pp. 128–142.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] T. Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, pp. 79–116, 1998.

[6] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[7] D. G. Lowe, "Object recognition from local scale-invariant features," pp. 1150–1157, 1999.

[8] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[9] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," pp. 404–417, 2006.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," pp. 886–893, 2005.

[11] M. Rusiñol, D. Karatzas, A. Bagdanov, and J. Lladós, "Multipage document retrieval by textual and visual representations," in *Proceedings of the International Conference on Pattern Recognition*, 2012, pp. 521–524.

[12] M. Rusiñol and J. Lladós, "Logo spotting by a bag-of-words approach for document categorization," in *10th International Conference on Document Analysis and Recognition*, 2009, pp. 111–115.

[13] G. David, "Distinctive image features from scale-invariant keypoints," pp. 91–110, 2004.

[14] B. Fulkerson, A. Vedaldi, and S. Soatto, "Localizing objects with smart dictionaries," in *European Conference on Computer Vision*, ser. Lecture Notes on Computer Science, 2008, vol. 5302, pp. 179–192.

[15] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117–128, January 2011.

[16] T. Sato, M. Iwamura, and K. Kise, "Fast and memory efficient approximate nearest neighbor search with distance estimation based on space indexing," IEICE Technical Report, Tech. Rep., Feb. 2013.

[17] T. Nakai, K. Kise, and M. Iwamura, "Camera-based document image retrieval as voting for partial signatures of projective invariants," in *IEEE 8th International Conference on Document Analysis and Recgonition*, 2005, pp. 379–383.

[18] 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," pp. 1054–1058, 2011.

[19] W. Sun and K. Kise, "Similar manga retrieval using visual vocabulary based on regions of interest," in *International Conference on Document Analysis and Recgonition*, 2011, pp. 1075–1079.

[20] F. Porikli and T. Kocak, "Fast distance transform computation using dual scan line propagation," in *Real-Time Image Processing*, 2007.

[21] L. Ikonen and P. Toivanen, "Distance and nearest neighbor transforms on gray-level surfaces," *Pattern Recognition Letters*, vol. 28, no. 5, pp. 604 – 612, 2007.

[22] T. Sato, M. Iwamura, and K. Kise, "Fast and memory efficient approximate nearest neighbor search with distance estimation based on space indexing," IEICE Technical Report, PRMU2012-142, Feb. 2013, in Japanese.

[23] A. Babenko and V. Lempitsky, "The inverted multi-index," 2012.