Delaunay triangulation-based features for camera-based document image retrieval system

Q.B. Dang*, M. Rusiñol†, M. Coustaty*, M.M. Luqman*, C.D. Tran‡ and J-M. Ogier*
* L3i Laboratory, University of La Rochelle, France.
† Computer Vision Center (CVC), Universitat Autònoma de Barcelona , Spain.
‡ College of Information and Communication Technology, Can Tho University, Vietnam.
e-mail: quoc_bao.dang@univ-lr.fr

***Abstract*—In this paper, we propose a new feature vector, named DElaunay TRIangulation-based Features (DETRIF), for real-time camera-based document image retrieval. DETRIF is computed based on the geometrical constraints from each pair of adjacency triangles in delaunay triangulation which is constructed from centroids of connected components. Besides, we employ a hashing-based indexing system in order to evaluate the performance of DETRIF and to compare it with other systems such as LLAH and SRIF. The experimentation is carried out on two datasets comprising of 400 heterogeneous-content complex linguistic map images (huge size, 9800 X 11768 pixels resolution) and 700 textual document images.**

***Keywords*—Camera-based Document Image Retrieval, Delaunay Triangulation, feature descriptors, indexing**

I. Introduction and Related work

In this digital age, the explosion of the number of portable digital imaging devices has created a tremendous opportunity for camera-based document image retrieval applications. Users can access a huge amount of content on the Internet and a big challenge is to propose some tools to link real documents to those captured with digital devices. For instance, some augmented reality tools appear to propose similar contents to the users by simply capturing an image with their smartphones or cameras [1], [2], [3].

Camera-based document image retrieval system takes as input a part or the whole page document acquired as a query by a digital camera, and retrieves a document image that includes the query [1], [2], [3]. This task creates challenging images for recognition, because captured images can be affected by uneven lighting, low resolution, motion blur and perspective distortion problems [4].

In last decade, several camera-based document image retrieval systems using local features for real-time indexing and retrieval have been proposed. One of the main advantages of local features is that they have been demonstrated to be distinctive, robust, and segmentation free [5], [6].

It can be seen from the block diagram of an example system in Fig. 1 which has two main phases for a camera-based document image retrieval system. These include the indexing phase and retrieval phase. Both of which share feature extraction step, which is comprised of keypoint detection and description. For feature extraction and indexing phase, we usually have to choose suitable features and an indexing method, respectively.

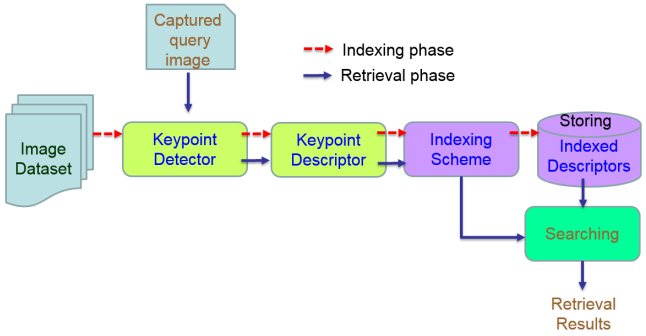

Fig. 1. Camera-based document image retrieval using local feature.

For local features, local keypoints are extracted in order to select parts of the image that will be retained for the description part. These local points and regions are generally capable of reproducing similar levels of performances to human observers; in locating elementary features in a wide range of image types. Local keypoint detectors are used to detect Regions of Interest (ROIs) that are invariant to a class of transformations (e.g. scaling, rotation and translation) so that for each detected region, which is usually represented as a keypoint, an invariant feature descriptor is built. Finally, these descriptors can be used as the basis to extract stable local image structures in a repeatable fashion and to encode them in a representation that is invariant to a range of image transformations, such as translation, rotation, scaling, and affine deformation [5], [6].

Recently, Rusinol et al. [7] built a system for spotting graphical symbols in camera-acquired documents in real time. They used ORB (Oriented FAST and Rotated BRIEF) [8] to extract features vectors and FLANN frame work [9] for indexing the features vectors, as well as for retrieving and spotting query images. According to the authors, ORB features are fast and efficient for real-time application. In this system, the database stores important information which includes symbols and logos. In the retrieval phase, these objects are recognized and spotted in the captured query.

In camera-based textual document image retrieval, the method called Locally Likely Arrangement Hashing (LLAH) is known as an efficient method with regard to accuracy, time and scalability [10], [11], [12], [13]. What is more important is that the authors proposed an efficient hashing technique, and LLAH has been shown superior to Geometric Hashing method concerning computational complexity [10], [14].

LLAH's feature extraction phase, as detailed in [10], [15], [16], [17], can be summarized as follows:

LLAH considers centroid of each word connected component (CC) as keypoints, which can be obtained even under perspective distortion, noise, and low resolution. A deep description on the method to obtain centroid of each word connected component can be found in [15]. From each keypoint P, the n nearest neighbor points around keypoint P are selected and organized clockwise. Then, all possible combination of m points among n are examined $(m < n)$. From one arrangement combination of m points, the LLAH vector is calculated based on a sequence of invariants calculated from all possible combinations of k points among m ($k = 3$ for similarity invariants, $k = 4$ for affine invariants and $k = 5$ for perspective invariants ; $k < m$).
When $k = 3$, the similarity invariant from 3 points A,B,C was defined as follows [17]:

$$\frac{AC}{AB} \tag{1}$$

When $k = 4$, the affine invariant from 4 points A,B,C,D was defined as follows [10]:

$$\frac{S(A,C,D)}{S(A,B,C)} \tag{2}$$

where S(A,B,C) is the area of a triangle with apexes A, B, C. When $k = 5$, the invariant of perspective transformation called cross-ratio from 5 points A,B,C,D,E was defined as follows [10]:

$$\frac{S(A,B,C)S(A,D,E)}{S(A,B,D)S(A,C,E)} \tag{3}$$

In order to reduce the sensibility of the system to keypoint extraction errors, multiple LLAH vectors are computed for each keypoint. As all the possible combinations of m points among n are examined, $\binom{m}{n}$ LLAH vectors have to be built from each keypoint. As a consequence, the more LLAH vectors are built, the more processing time and memory consumption the system requires. Thus, n and m need to be suitably set depending on each system.

LLAH using perspective and affine invariant works well, when the captured query is a complete document. Aiming to deal with portions of document captured by camera, Takeda *et al.* [11] proposed an extension of the LLAH feature by adding some additional features which are based on the rank of k area ratios of the extracted word regions. In another work, they also proposed to improve the LLAH features by adding additional features based on rank of areas of words regions [13]. Similarly, Kise *et al.* [12] improved the LLAH feature by using the rank of k areas of letter regions and the query expansion method in order to cope with small document portions captured by camera-pen system [12].

Recently, we have proposed the new feature vector, named Scale and Rotation Invariant Features (SRIF) for camera-based document image retrieval system [18]. SRIF is computed based on geometrical constraints between pairs of nearest points around a keypoint (as illustrated in Fig. 2). It can deal with feature point extraction errors which are introduced as a result of the camera capturing of documents. From each pairs of points around P, SRIF combines the scale and rotation invariant which is $\theta_{ij}\ max(|\overrightarrow{PP_i}|/|\overrightarrow{PP_j}|, |\overrightarrow{PP_j}|/|\overrightarrow{PP_i}|)$. SRIF works well, even when the captured query represents only a small portion of a document.

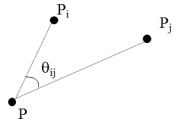

Fig. 2. Constraints between two point around one keypoint P.

When using LLAH and SRIF, two vital parameters n and m that control combination of local keypoints and compute descriptors need to be set. We aim to propose a new descriptor which can be employed without parameters controlling selection of feature points. Our idea is to use a stable structure of the feature points and then to build descriptors from this structure so that it can scope with portions of document captured by camera. Because of this, we choose Delaunay triangulation to form the stable structure for the feature points.

Delaunay triangulation has three main properties [19]:

- Given a set of points, there always exists a Delaunay triangulation except when all the points are aligned.
- The Delaunay triangulation maximize the minimum angle of each triangle in the triangulation.
- When a subset of four or more points can be placed on the same circle (e.g. the vertices of a rectangle), the Delaunay triangulation of the points is not unique.

From these properties, we will always be able to compute a Delaunay triangulation from centroids of word CCs in documents. In the case of instability, aligned points, it will never occur in the whole page, it can occur locally and create local instabilities. To deal with this, we propose an efficient way to combine the local Delaunay triangulation in order to build efficient descriptors.

In this paper, we propose a new feature vector, named DETRIF. DETRIF is computed based on the geometrical constraints from the Delaunay triangulation which is constructed from centroids of connected components. Our main contributions is this efficient DETRIF features for camera-based document retrieval.
The rest of this paper is organized as follows. In Section II, we present details about how DETRIF descriptors are built, indexed and retrieved. Section III presents how datasets and ground truths are built and the experimental results. Finally, the conclusion and future work are given in Section IV.

II. Delaunay Triangulation Feature Extraction

In this section, we will present how DETRIF (standing for DElaunay TRIangulation-based Features) descriptors are extracted, indexed and retrieved. As the Delaunay triangulation is invariant to similarity transformations and not to perspective, invariant values of DETRIF are extracted from geometrical

constraints on each pair of adjacency triangles in Delaunay triangulation. Therefor, this feature is tolerant to perspective distortion [20], [21].

A. Feature extraction

DETRIF considers centroids of word connected components as feature points from which Delaunay triangulation is constructed. An example of DETRIF computation on a map is illustrated in Fig. 3.

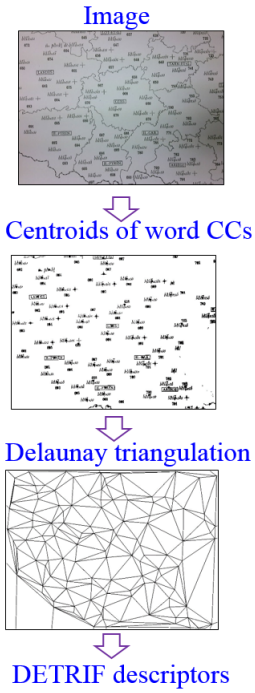

Fig. 3. The main steps to build DETRIF descriptors.

The pseudo-code below describes how DETRIF descriptors are built (for a document image), from a Delaunay triangulation.

For each triangle tr ***from Delaunay triangulation do***
For each vertex v ***in*** tr ***do***
If exist adjacency triangle of tr ***then //(Fig. 4)***
For each vertex X_i ***connect to*** v ***and not belong to*** tr ***do***
Build DETRIF descriptor f ***at*** X_i ***//(Fig. 5)***
Compute H_{index} ***using Equation (5)***
Store f***(Document ID, Point ID) in hash table by using*** H_{index}
End for
End if
End for
End for

In order to reduce the sensibility of the system to keypoint extraction errors, DETRIF descriptor is built for each vertex X_i. As we can see in Fig. 5. The constraint between X_i and two triangles $\triangle ABD$ and $\triangle BCD$ are used to build one DETRIF descriptor. This process can be summarized in two steps:

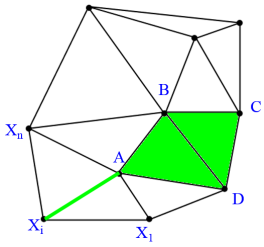

Fig. 4. Adjacency triangle (ABD, BDC) and vertexs connected from vertex A ($X_0, X_1, ...X_n$).

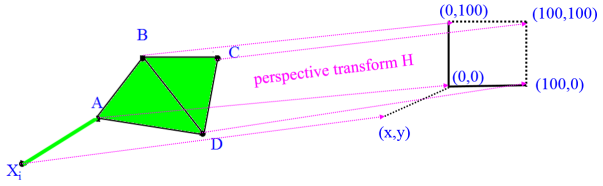

Fig. 5. DETRIF descriptor extraction from each vertex X_i

Firstly, we find the perspective transformation H that transforms four points $ABCD$ on the left of Fig. 5 into a normalized coordinate system on the right of Fig. 5. Then, transformation H is applied on X_i in order to obtain the (x,y) coordinates in the new normalized space. Aiming to get positive features for indexing DETRIF descriptors with a hash table, the point (x,y) is transformed into polar coordinate system. As a result, the point (x,y) becomes the point (r,φ) in polar coordinate system, where $r \in R$, and $\varphi \in (0..360^o)$. These two invariant values are used to build DETRIF descriptor. In order to make DETRIF descriptor more distinctive we also use geometric constraints from Fig. 5. Finally, DETRIF descriptors f are built by using the invariant values that includes r, φ, $\frac{S(B,C,D)}{S(A,B,D)}$, $\angle ABC$, $\angle BCD$, $\angle CDA$, $\angle DAB$, $\angle X_iAB$, $\angle X_iBC$, $\angle X_iCD$, $\angle X_iAD$. Where, $\angle ABC$ is the angle between $\overrightarrow{AB}$ and $\overrightarrow{BC}$, so $\angle ABC \in (0..\pi)$

B. Indexing phase

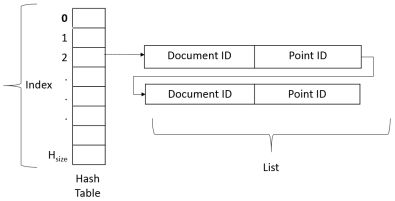

Fig. 6. The hash table structure.

To index DETRIF descritors, we employ a hashing-based indexing system which is similar to the method presented in [15] (Fig. 6 presents the hashing strategy). DETRIF vectors (called f) can be indexed and retrieved very quickly using one hash table even if they are not stored in the hash table for checking distances of nearest neighbors, which speeds up the

retrieval time and reduces the amount of used memory. Furthermore, this indexing scheme allows adding new documents into database without rebuilding all the database structure of indexes.

These performances rely on the use of integer feature vectors f, that are discretized and normalized as follows:

$$f(i) = trunc(f(i)) * 2 + round(f(i) - trunc(f(i))) \quad (4)$$

This normalization makes angles added to DETRIF vectors be more tolerant to perspective distortions.

The hash function that is used for indexing DETRIF vectors is defined as follows:

$$H_{index} = (\sum_{i=0}^{d-1} f_i q^i) \bmod H_{size} \quad (5)$$

where d is the number of dimensions of vector f, q is the level of quantization constant (e.g. $q = 17$), H_{size} is the size of hash table.

In order to add a new document into database, the system firstly extracts keypoints from centroids of word connected components. Then for each keypoint, all DETRIF vectors are computed and indexed. As shown in Fig. 1, both indexing and retrieval share the feature extraction and use the same hash function (5).

C. Retrieval phase

Starting from a query image captured with a camera, keypoints are firstly extracted. Then DETRIF vectors are computed like in features extraction (II-A). Then it is looked up in the indexing system in order to get the list of document IDs related to each keypoint (Fig. 6). The pseudo-code is described as below.

For each triangle tr ***from Delaunay triangulation do***
For each vertex v ***in*** tr ***do***
If exist adjacency triangle of tr ***then //(Fig. 4)***
For each vertex X_i ***connect to*** v ***and not belong to*** tr ***do***
Build DETRIF descriptor f ***at*** X_i ***//(Fig. 5)***
Compute H_{index} ***using Equation (5)***
Look-up in hash table by using H_{index}
and do voting and validating
End for
End if
End for
End for

For each document in the retrieval result list, the number of votes for it in the voting table is incremented. After getting the voting result, the top-t documents with largest number of votes are selected as candidate results.

In order to check the correctly matched results in top-t returned documents. For each document in top-t, it must be ensured that whether or not there is a correct perspective transformation between query's keypoints and matched document's keypoints. To validate this, RANSAC [22] is used. If no best transformation can be found, the number of votes of document is set to zero. Lastly, the document with majority of votes in top-t result documents is returned as the result. A correct retrieval result is validated if it has a correct document ID on one hand, and if it corresponds to the correct region of the document on the other hand.

To validate the correct region, first RANSAC is applied so that we can obtain the spotting region of query image in the returned document through perspective transformation. Next, the overlap between the ground-truth region (where query image was captured) and the spotted region is computed. The frame is considered as a correct retrieval result if the area of the overlap is more than 60 percent of the area of the spotting region otherwise it is considered as an incorrect result. An example of the overlap region validation is shown in Fig. 7.

III. Experimentation

In this section first we describe the dataset and its ground truth generation. Afterwards we present the details on the experimentation and a detailed discussion on the obtained results.

A. Dataset and the ground truth generation

To validate DETRIF as well as to compare it with LLAH and SRIF, we evaluated both retrieval and spotting accuracies. Moreover, retrieval time is also considered. The experiment is tested with two datasets. The wikibook dataset represents the images with textual content only. The cartodialect dataset represents images with graphical content mainly.

For the WikiBook dataset, we chose a book from wikibooks[1] including 700 A4-sized pages, which are scanned at the resolution of 300 dpi in JPEG format.

The CartoDialect dataset includes French linguistic maps, and is composed of 400 images with a resolution of 9800 x 11768 pixels. Each map contains the phonetic symbols which describe the pronunciation of a word in different regions of France. All maps contain the same graphical elements which are region borders. Moreover, text density in each map is very sparse.

In order to build the ground truth for each data set, the document in WikiBook dataset is divided into 4 regions (top left, top right, bottom left and bottom right). Because the size of document in CartoDialect dataset is large, each document is divided into 6 regions (top left, top right, middle left, middle right, bottom left and bottom right, see Fig. 7 for details. The information of region is also used for validating the correct spotting in retrieval phase by dividing the database images into 6 regions with the same way.

One video was recorded at each region except blank regions. Documents were captured without rotations. The IPEVO VZ-1 HD document camera was used for recording the videos. It was fixed at 10 to 15cm above surface of the captured document, and the resolution of the captured images was 1024x768.

For each video, we selected the first 15 frames. To validate the rotation invariance, we also rotated each frame by an angle of 0, 90, and 180 degrees. We choose two specific angle because it does not affect too much the keypoints which were extracted by

[1] http://upload.wikimedia.org/wikipedia/commons/2/2d/LaTeX.pdf

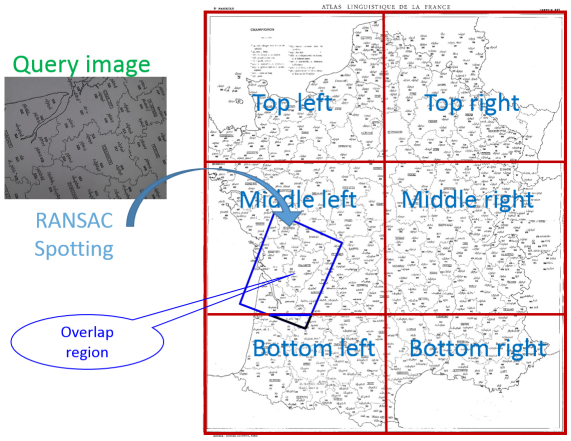

Fig. 7. Captured video from a map at six regions, the overlap between spotting region results and captured region from a query image.

a connected component(CC) extraction algorithm. The number of captured videos is shown in Table I. These datasets and their ground truth are publicly available for academic research purposes[2].

TABLE I. DATASETS DETAILS

Dataset name	# of documents	Resolution	# of videos	# of tested frames
WikiBook	700	2480 x 3508	1630	24450
CartoDialect	400	9800 x 11768	2400	36000

B. Experimental protocol and the evaluation measure

In order to evaluate all methods (DETRIF, LLAH and SRIF) we measured the retrieval accuracy and the average retrieval time per each frame. For each video, we evaluated the retrieval accuracy called as the *video retrieval accuracy*. For this evaluation, 15 frames were extracted from each video, and each frame was rotated by an angle of 0^o, 90^o, or 180^o before going to the retrieval phase. If number of correctly retrieved frames are greater than 50% of total frames (15 frames) extracted from the video, video was considered as successful. Otherwise video was considered as failed. This threshold ensures that it is the majority returned result. Finally, videos retrieval accuracy is the ratio between the number of correct retrieval videos and the total of videos ground truth from each dataset.

LLAH, SRIF and DETRIF shared the same keypoint extraction approach which is based on the extraction of centroids of word connected components and share the indexing framework. We also used the method in [23] to discard the borders in the maps and to extract centroids of word CCs. Besides, small CCs which are noise were discarded.

All methods use the same voting and validating in retrieval phased. LLAH was tested with three invariants that are affine (LLAH-Affine), perspective (LLAH-Perspective) and similarity (LLAH-Similarity) invariant . For SRIF, LLAH-Affine, and LLAH-Similarity we set $n = 7$, $m = 6$ (it got better than results with $n = 6$, $m = 5$) without adding additional features. For LLAH-Perspective n and m are set equaling to 8,7 respectively. $H_{size} = 10^{17}$, $t = 10$ for selecting top-t of best candidate retrieval results. To avoid collisions in the hash table wet set $q = 15$ for LLAH Affine, $q = 3$ for LLAH Similarity, $q = 2$ for LLAH-Perspective and $q = 37$ for DETRIF. Our systems were implemented on a 64 GB RAM Linux machine running in C extended C++ environment with a single thread.

[2]For academic research purposes the dataset can be downloaded from http://navidomass.univ-lr.fr/SRIFDataset/

C. Experimental results

1) WikiBook dataset:

Testing results on this dataset are shown in Table II. As it can be seen that retrieval time of all methods is more or less 0.3 second/query, and the fasted method is SRIF. SRIF and LLAH-Similarity got the highest retrieval accuracy with around 84%. Concerning retrieval accuracy, DETRIF is the third with 74.1% which is higher than LLAH-Affine and LLAH-Perspective.

TABLE II. THE TESTING RESULTS ON WIKIBOOK DATASET

Method	Videos Retrieval Accuracy				Retrieval Time (s/query)			
	0^o	90^o	180^o	Avg	0^o	90^o	180^o	Avg
DETRIF	77.5%	73.4%	71.5%	74.1%	0.36	0.37	0.38	0.37
SRIF	84.1%	84.2%	84.3%	**84.2%**	0.32	0.32	0.33	**0.32**
LLAH-Similarity	83.9%	84.4%	84.1%	84.1%	0.32	0.33	0.33	0.33
LLAH-Affine	53.8%	52.6%	51.9%	52.8%	0.34	0.34	0.34	0.34
LLAH-Perspective	56.3%	55.8%	54.9%	55.7%	.38	0.39	0.39	0.39

2) CartoDialect dataset:

TABLE III. THE TESTING RESULTS ON CARTODIALECT DATASET

Method	Videos Retrieval Accuracy				Retrieval Time (s/query)			
	0^o	90^o	180^o	Avg	0^o	90^o	180^o	Avg
DETRIF	95.8%	94.2%	93.6%	94.5%	0.35	0.38	0.41	0.38
SRIF	95.7%	95.4%	94.6%	**95.2%**	0.28	0.28	0.29	**0.28**
LLAH-Similarity	95.0%	94.5%	93.9%	94.5%	0.38	0.37	0.39	0.38
LLAH-Affine	81.7%	80.8%	80.5%	81.0%	0.64	0.62	0.59	0.62
LLAH-Perspective	16.5%	15.7%	15.2%	15.8%	1.0	1.2	1.1	1.1

The experimental results on CartoDialect dataset are shown in Table III. The best performance methods is SRIF in terms of accuracy retrieval and retrieval time with 95.2% and 0.28 second/query respectively. DETRIF and LLAH-Similarity are the second best performance methods. Both of them got 94.5% accuracy retrieval with 0.38 second/query, which is approximately SRIF's result. LLAH-Affine got a lower accuracy retrieval with 81.0% and its retrieval time is also slower than retrieval time of DETRIF as well as LLAH-Similarity. LLAH-Perspective got the lowest accuracy retrieval and the slowest retrieval time.

D. Discussion

Although DETRIF needs more time to build Delaunay triangulation structure compared to LLAH and SRIF, it still got a fast retrieval time when number of feature points is not too large in each query. Because the Delaunay triangulation of a set S of N points in the plane can be computed in $O(NlogN)$ expected time.

Let S be a set of N points in the plane, not all collinear, and let K denote the number of points in S that lie on the boundary of the convex hull of S. Then any triangulation of P has $2N-2-K$ triangles and $3N-3-K$ edges. The computational complexity of building DETRIF descriptors depend on the

time to find the adjacency vertexes and triangle. So, the computational complexity of building DETRIF descriptors is $O(N)$ which is similar to the computational complexity of LLAH.

The reason why retrieval accuracy of all methods was not so good in WikiBook dataset is that the number of word CCs is insufficient in many queries (as show in Fig. 8).

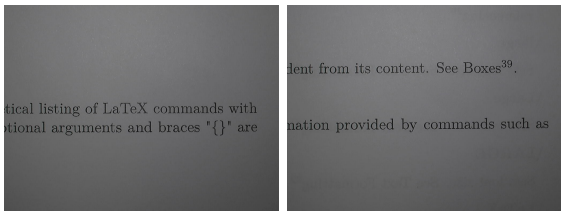

Fig. 8. Insufficient word CCs query examples in WikiBook dataset.

IV. Conclusion

We have presented the new features namely DETRIF which are built from the Delaunay triangulation of features points. DELTRF is the paramter-less method compared with LLAH and SRIF and give a high performance. In addition, we built a dataset and the ground truth that is composed of 400 very-large heterogeneous-content complex linguistic map images. The experimental results in this dataset show that DELTRF can correctly deal with the context of documents containing small numbers of texts; furthermore DETRIF outperformed LLAH using affine invariant and LLAH using perspective invariant and. DETRIF got equivalent performance with LLAH using similarity invariant from both the retrieval accuracy point of view, and processing time point of view.

In the future, we are going to improve our new feature and evaluate it on other datasets with perspective distortion. Besides, we are going to try to use DETRIF with dedicated keypoints such as SUFR, SIFT, ORB or FAST in order to investigate into generic descriptors for information spotting in huge repositories of scanned document images.

Acknowledgment

This work has been partially supported by the ECLATS project which is funded by the French National Research Agency under grant ANR-15-CE38-0002-01, and by the Spanish Ministry of Education and Science under projects TIN2014-52072-P, and TIN2012-37475- C02-02, by the People Programme (Marie Curie Actions) of the Seventh Framework Programme of the European Union (FP7/2007-2013) under REA grant agreement no. 600388, and by the Agency of Competitiveness for Companies of the Government of Catalonia, ACCIO.

References

[1] Q. Liu and C. Liao, "Paperui," in *Camera-Based Document Analysis and Recognition*. Springer, 2012, pp. 83–100.

[2] K. Takeda, K. Kise, and M. Iwamura, "Real-time document image retrieval on a smartphone," in *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*. IEEE, 2012, pp. 225–229.

[3] J. J. Hull, B. Erol, J. Graham, Q. Ke, H. Kishi, J. Moraleda, and D. G. Van Olst, "Paper-based augmented reality," in *Artificial Reality and Telexistence, 17th International Conference on*. IEEE, 2007, pp. 205–209.

[4] J. Liang, D. Doermann, and H. Li, "Camera-based analysis of text and documents: a survey," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 7, no. 2-3, pp. 84–104, 2005.

[5] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and Trends® in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[6] J. Li and N. M. Allinson, "A comprehensive review of current local features for computer vision," *Neurocomputing*, vol. 71, no. 10, pp. 1771–1787, 2008.

[7] M. Rusinol, D. Karatzas, and J. Lladós, "Spotting graphical symbols in camera-acquired documents in real time," in *Graphics Recognition. Current Trends and Challenges*. Springer, 2014, pp. 3–10.

[8] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: an efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2564–2571.

[9] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration." *VISAPP*, vol. 2, 2009.

[10] T. Nakai, K. Kise, and M. Iwamura, "Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval," in *Document Analysis Systems VII*. Springer, 2006, pp. 541–552.

[11] 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 *2011 International Conference on Document Analysis and Recognition*, Sep. 2011, pp. 1054–1058.

[12] K. Kise, M. Chikano, K. Iwata, M. Iwamura, S. Uchida, and S. Omachi, "Expansion of queries and databases for improving the retrieval accuracy of document portions: an application to a camera-pen system," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*. ACM, 2010, pp. 309–316.

[13] T. Nakai, K. Kise, and M. Iwamura, "Real-time retrieval for images of documents in various languages using a web camera," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*. IEEE, 2009, pp. 146–150.

[14] H. J. Wolfson and I. Rigoutsos, "Geometric hashing: An overview," *Computing in Science and Engineering*, vol. 4, no. 4, pp. 10–21, 1997.

[15] T. Nakai, K. Kise, and M. Iwamura, "Camera based document image retrieval with more time and memory efficient llah," *Proc. CBDAR*, pp. 21–28, 2007.

[16] ——, "Hashing with local combinations of feature points and its application to camera-based document image retrieval," *Proc. CBDAR05*, pp. 87–94, 2005.

[17] M. Iwamura, T. Nakai, and K. Kise, "Improvement of retrieval speed and required amount of memory for geometric hashing by combining local invariants," in *Proc. 18th British Machine Vision Conference (BMVC2007)*, vol. 2, Sep. 2007, pp. 1010–1019.

[18] Q. Dang, M. Luqman, N. Coustaty, M, C. Tran, and J. Ogier, "Srif: Scale and rotation invariant features for camera-based document image retrieval," in *Document Analysis and Recognition, 2015. ICDAR'15. 13th International Conference on*. IEEE, 2015, pp. 601–605.

[19] D.-T. Lee and A. K. Lin, "Generalized delaunay triangulation for planar graphs," *Discrete & Computational Geometry*, vol. 1, no. 1, pp. 201–217, 1986.

[20] G. D. Evangelidis and C. Bauckhage, "Efficient subframe video alignment using short descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 10, pp. 2371–2386, 2013.

[21] P. McIlroy, S. Izadi, and A. Fitzgibbon, "Kinectrack: Agile 6-dof tracking using a projected dot pattern," in *Mixed and Augmented Reality (ISMAR), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 23–29.

[22] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[23] Q. Dang, M. Luqman, M. Coustaty, N. Nayef, C. Tran, and J. Ogier, "A multi-layer approach for camera-based complex map image retrieval and spotting system," in *Image Processing Theory, Tools and Applications (IPTA), 2014 4th International Conference on*. IEEE, 2014, pp. 1–6.