Quality assurance for document image collections in digital preservation ⋆

Reinhold Huber-Mörk[1] and Alexander Schindler[1,2]

[1] Research Area Intelligent Vision Systems
Department Safety & Security
Austrian Institute of Technology
reinhold.huber-moerk@ait.ac.at
[2] Department of Software Technology and Interactive Systems
Vienna University of Technology
schindler@ifs.tuwien.ac.at

Abstract. Maintenance of digital image libraries requires to frequently asses the quality of the images to engage preservation measures if necessary. We present an approach to image based quality assurance for digital image collections based on local descriptor matching. We use spatially distinctive local keypoints of contrast enhanced images and robust symmetric descriptor matching to calculate affine transformations for image registration. Structural similarity of aligned images is used for quality assessment. The results show, that our approach can efficiently asses the quality of digitized documents including images of blank paper.

1 Introduction

Large collections of image data include scanned or rendered document image data from historical archives or large-scale document preservation activities such as digital museum collections or the Google books initiative⋆⋆. It is commonly observed that different versions of image collections with identical or near-identical content exist in such collections resulting from independent acquisitions or repeated downloads of Google books image collections with different post-processing, e.g. rectification, denoising, compression, rescaling, cropping etc.

We describe an approach for analysis and comparison of collections of digital document image data. Maintainers of image archives, such as libraries, as well as researchers in the field of digital preservation are typically confronted with inconsistencies in their collections. Due to independent acquisitions, digital file format migration or modification of image properties the task of content verification arises. Additionally, the long-term storage of data using deprecated hardware and data formats results in issues such as bit rot or limited or difficult access to the data.

From the point of document image content comparison a robust approach is required. Recently, stable image feature descriptors invariant to geometric and radiomet-

⋆ This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137)

⋆⋆ http://www.google.com/googlebooks/library.html

ric distortions became state of the art for various applications in computer vision. Although the obtained quality and plausibility for human operators meets high expectations in various domains, image feature extraction and comparison based on advanced image analysis methods requires significant computational resources. Furthermore, image data in large collections of scanned documents is characterized by large volumes of data, therefore a compact representation of content information is aspired. Content-based representation and comparison of images requires image rectification and the use of expressive measures of structural similarity. Stable image correspondence is based on matching of local descriptors. From the point of storage consumption, the use of a condensed representation based on local descriptors enables a reduction of data volumes.

This paper is organized as follows. Section 2 provides an overview of document image comparison as well as the application of local features. In Sect. 3 we present our approach to document comparison. Section 4 shortly discusses image quality assessment as used in our work. Results are presented and Sect. 6 and Sect. 7 summarizes our work.

2 Related Work

Related work in the field of analysis of document image collections include tasks such as indexing, revision detection, duplicate and near-duplicate detection. Several authors mention that the use of optical character recognition, which is an obvious approach to extract relevant information from text documents, is quite limited with respect to accuracy and flexibility [2, 7, 17]. An approach combining page segmentation and Optical Character Recognition (OCR) for newspaper digitization, indexing and search was described recently [5], where a moderate overall OCR accuracy on the order of magnitude of 80 percent was reported. Page Segmentation is prerequisite for the document image retrieval approach suggested in [2] where document matching is based on the earth mover's distance measured between layout blocks. The PaperDiff system [17] finds text differences between document images by processing small image blocks which typically correspond to words. PaperDiff can deal with reformatting of documents but is restricted as it is not able to deal with documents with mixed content such as pages containing images, blank pages or graphical art. A method for duplicate detection in scanned documents based on shape descriptions for single characters also showed advantages with respect to robustness and speed when compared to OCR [7]. The most similar work, compared to our paper, is a revision detection approach for printed historical documents [3]. Contrarily to our approach, connected components are extracted from document images and Recognition using Adaptive Subdivisions of Transformation (RAST) [4] was applied to overlay images and highlight differences without providing details on the comparison strategy.

Apart from document image processing, several approaches to duplicate and near-duplicate image detection and image and sub-image retrieval were published in the related field of video and web image processing. Typically, approaches in this area make use of local image descriptors to match or index visual information. Near-duplicate detection of keyframes using one-to-one matching of local descriptors was described

for video data [26]. A bag of visual keywords [6], derived from local descriptors, was described as an efficient approach to near-duplicate video keyframe retrieval[23]. For detection of near-duplicates in images and sub-images local descriptors were also employed [12].

In general, the application of local features ranges from texture recognition, robot localization to wide baseline stereo matching and object class recognition. In spite of their success and generality, these approaches are limited by the distinctiveness of the features and the difficulty of appropriate matching [8]. A survey and evaluation on the performance of local features in the context of their repeatability in the presence of rotation, scale, illumination, blur and viewpoint changes is provided in [14]. One of the most prominent local keypoint detection and description method, the Scale Invariant Feature Transform (SIFT) [13] descriptor is based on gradient distribution in salient regions. Faster keypoint detection and description method include Features from Accelerated Segment Test (FAST) [18], Speeded up Robust features (SURF) [1], and the recently developed Oriented Brief (ORB) based on Binary Robust Independent Elementary Features (BRIEF) [19].

3 Document image processing

Pixel-wise comparison of images is only possible as long no geometric modifications were applied. Additionally, in cases of filtering, color or tone modifications the information at the pixel level might differ significantly, although the image content is well preserved. Therefore, we suggest to use interest point detection and derivation of local feature descriptors, which have proven highly invariant to geometrical and radiometrical distortions [13, 22] and were successful applied to a variety of problems in computer vision.

To detect and describe interest regions in document images we used the SIFT keypoint extraction and description approach. SIFT selects an orientation by determining the peak of the histogram of local image gradient orientations at each keypoint location. Subpixel image location, scale and orientation are associated with each SIFT descriptor (4×4 location grid $\times$ 8 gradient orientations). The keypoint locations itself are identified from a scale space image representation. All keypoints with low contrast or keypoints that are localized at edges are eliminated using a Laplacian function.

3.1 Contrast enhancement

When investigating collections of historical artifacts it turned out that images of blank pages, e.g. images with no text or graphics are also important for historians. Therefore, images of blank pages need to be considered as well as images containing graphical art. In order to treat images with high textual or graphical information content as well as images showing blank pages we adopted a procedure for local contrast enhancement called Contrast-Limited Adaptive Histogram Equalization (CLAHE) [16]. Fig. 1 shows image pairs before and after local contrast enhancement. Clearly, the paper structure on blank images is enhanced while in region of rich information only small modifications are observed. Figure 1 shows two images before, one blank image and one image

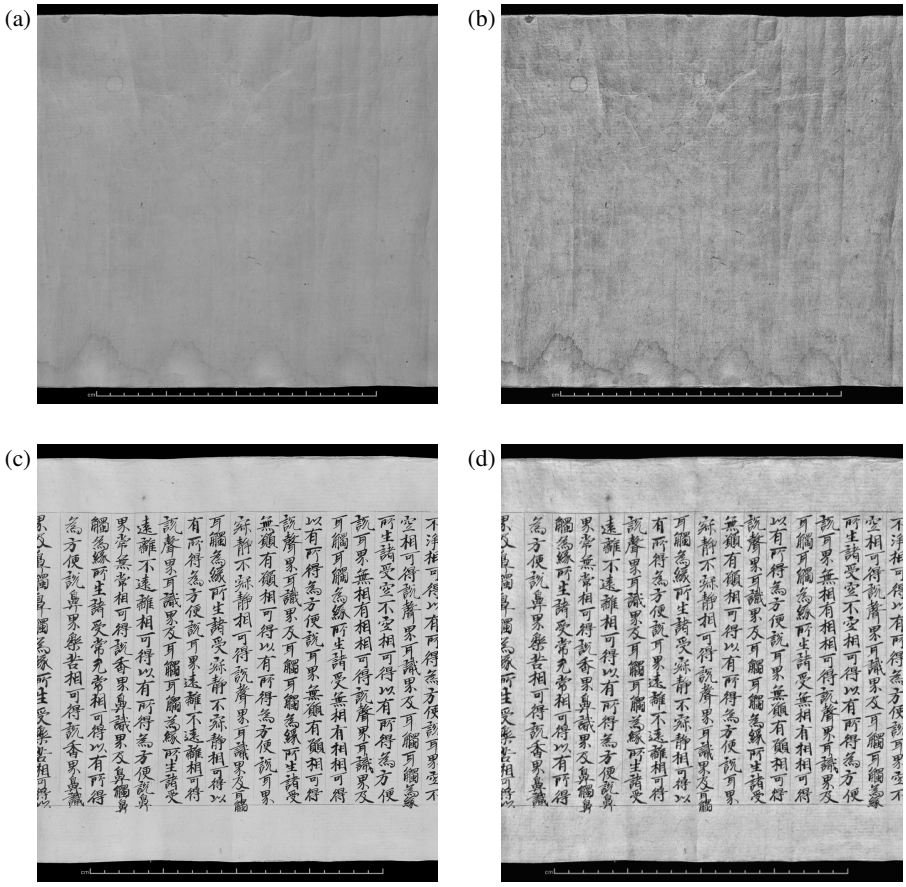

Fig. 1. Contrast enhancement of images: (a) blank image original, (b) blank image after CLAHE, (c) text image original, (d) text image after CLAHE

containing text, and after application of CLAHE. Especially in blank page the paper structure is enhanced while in the text image lesser tone modifications are observed.

3.2 Robust symmetric matching

As suggested by Lowe [13], local descriptors are matched by identifying the first two nearest neighbors in Euclidean space. A descriptor is accepted only if the distance ratio to the second nearest neighbor is below a given threshold. An essential characteristic of this approach is that a descriptor can have several matches when different descriptors from the second image matched against the same descriptor from the first image. The overcome this problem, one can either ignore all ambiguous matches or keep the one with lowest distance. We also adopted this idea by enforcing one-to-one matching

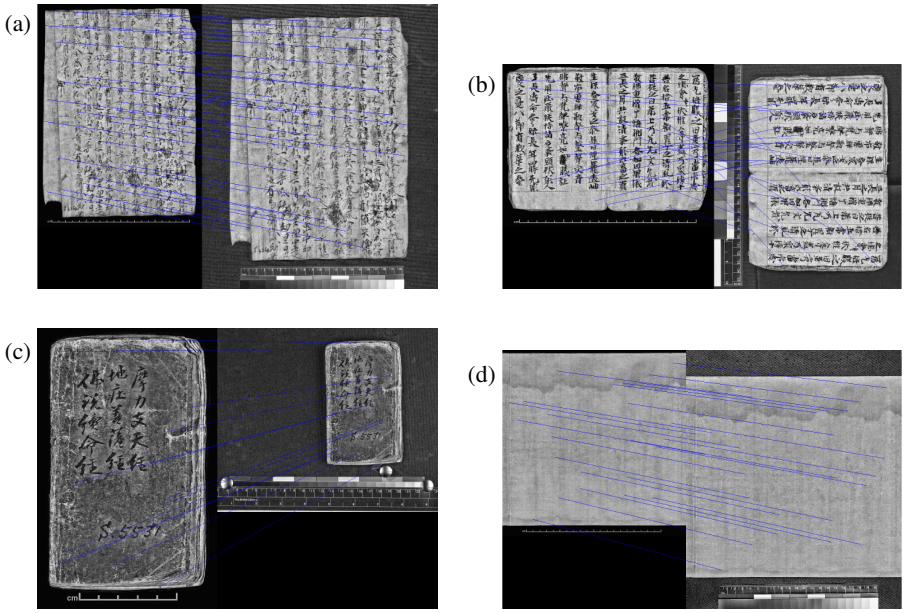

Fig. 2. Examples for matching of spatially distinctive local keypoints: (a) image pair containing text, (b) rotated image pair, (c) image pair with dirt and large scale and content difference, (d) blank image pair.

of descriptors. Figure 2 shows different typical cases of image pairs, e.g. text, rotated, noisy and blank images, with obtained correspondences of keypoints between the images overlaid as lines.

3.3 Spatially distinctive local keypoints

Spatially distinctive keypoints are derived from local interest regions. The approach of local interest regions was inspired by the spatially aligned pyramid matching [24] where images are divided into rectangular overlapped and non-overlapped image blocks. In document images, scale variation is limited and it is not necessary to employ pyramid or scale-space schemes for the selection of interest regions. On the other hand, for the task of keypoint extraction a scale-space representation, as inherent to SIFT descriptors, is very valuable.

Centers of local interest regions are simply formed by a regular grid overlaid on the image with grid positions given by

$$(u_{i,j}, v_{i,j}), \quad 1 \leq i \leq m, 1 \leq j \leq n, \tag{1}$$

where m an n denote the number of grid points in horizontal and vertical dimensions. The overall number of interest regions is $k = m \cdot n$. The region of influence for each

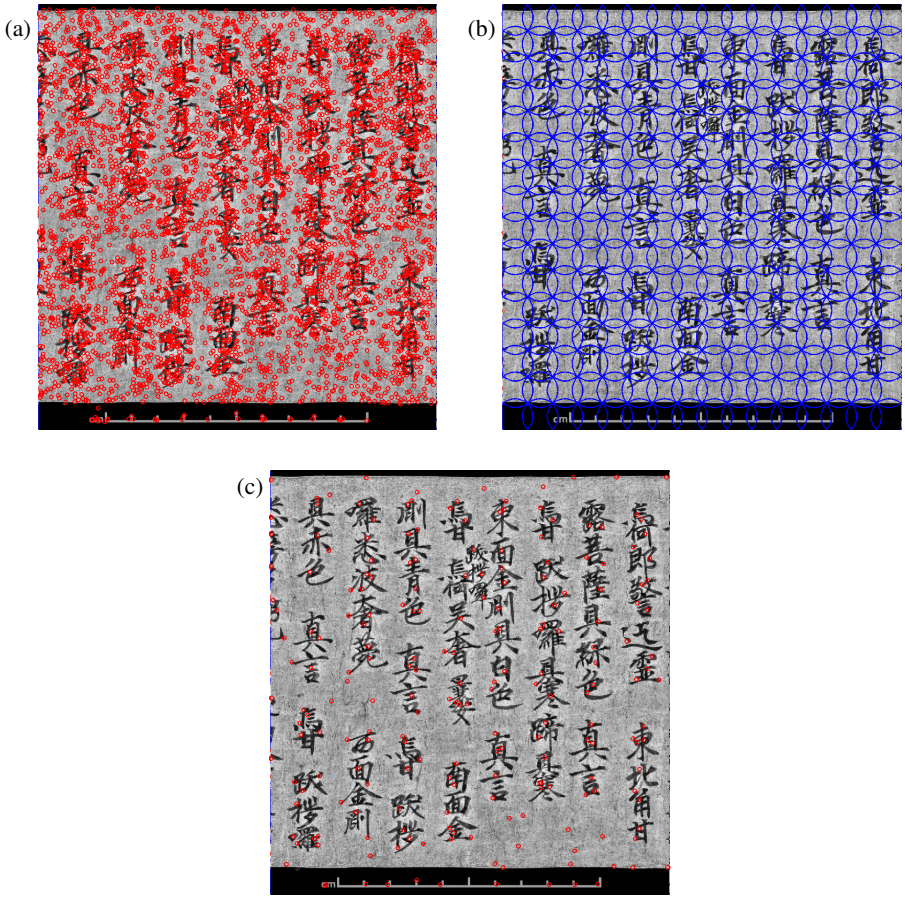

Fig. 3. Selection of spatially distinctive local keypoints: (a) image with all keypoints, (b) local interest regions, (c) spatially distinctive keypoints.

center local region of interest is described by a region of circular shape. The distance between given grid centers is given by $d = M/m = N/n$ and the influence area for each interest region becomes $r = \sqrt{2} \cdot d$. Image dimensions are $M \times N$, where M is the horizontal and N is the vertical resolution. As grid spacing d is commonly not an integer valued number, it has to be ensured to cover the full image domain. Figure 3 (a) shows all keypoints found in an image and Fig. 3 (b) the overlaid, partially overlapping interest regions.

For each interest region centered at $(u_{i,j}, v_{i,j})$ we search for the keypoint with highest saliency in the circular neighbourhood given by the search radius R. We employed the Harris corner detection approach [11] as simple measure of saliency in oder to select keypoints inside interest regions. The 2D structure tensor A for pixel position (x, y) is

given by

$$A(x,y) = \begin{bmatrix} I_x^2(x,y) & I_x(x,y)I_y(x,y) \\ I_x(x,y)I_y(x,y) & I_y^2(x,y) \end{bmatrix}, \tag{2}$$

where I_x and I_y are the partial derivatives of the image with respect to directions given by x and y. The eigenvalues of A provide information about the local image structure. Corner points are regarded as stable points for which both eigenvalues are large. In order to avoid eigenvalue decomposition, the following measure of corner strength was suggested by Harris and Stephens [11]

$$R(x,y) = \det A(x,y) - k \,\text{trace}^2 A(x,y), \tag{3}$$

where k is s constant, commonly chosen as $k = 0.04$. The set of spatially distinctive keypoints is derived from the strongest corner points from each local interest regions. Figure 3 (c) shows the selected keypoints with respect ro the interest regions.

3.4 Descriptor matching

The matching of spatially distinctive keypoint descriptors is based on the established robust matching method called Random Sample Consensus (RANSAC) [9], where corresponding points are randomly drawn from the set of spatially distinctive keypoints and the concensus test is constrained on an affine fundamental matrix describing the transformation between image pairs. The obtained affine transformation parameters are used to overlay corresponding images by warping one image to the other.

4 Quality assessment

Image quality assessment can roughly be dived into reference-based (non-blind) [20, 22, 25] and no reference-based (blind) [10, 15] evaluation. Intermediate definitions exist, but they are of minor interest in the context of our paper. Blind image quality assessment considers single images and tries to quantify their information content either based on low level image features or using elaborate machine learning techniques. The setup we are dealing with is to find out severe differences in content, which is addressed by non-blind image quality assessment. In such a setup, differences of visual appearance are quantified and the decision which image in a pair of images is visually more appealing is left to the human observer.

It is well known that image difference measures such as taking the mean squared pixel difference does not correspond to the human perception of image difference [21]. Therefore, we employed the structural similarity image (SSIM) non-blind quality assessment [22]. SSIM basically considers luminance, contrast and structure terms to provide a measure of similarity for overlaid images. The SSIM $s(I_1(x,y), I_2(x,y))$ between images I_1 and I_2 is calculated at each pixel location (x,y). In order to correct for small errors in image rectification we calculated the local minima $s_{\mathcal{N}}(x,y)$ of the SSIM between $I_1(x,y)$ and $I_2(\mathcal{N}(x,y))$, i.e. between the image I_1 and shifted versions

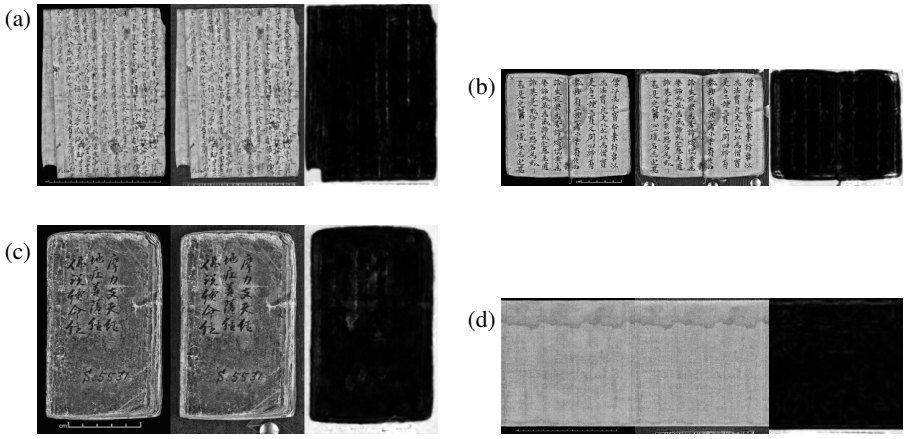

Fig. 4. Examples of image pairs after rectification and structural similarity at each pixel (black ... high structural similarity, white ... low structural similarity): (a) image pair containing text, (b) rotated image pair, (c) image pair with dirt and large scale and content difference, (d) blank image pair..

of I_2. For the neighborhood $\mathcal{N}$ we used shifts of one pixel into all eight adjacent pixel directions

$$s_{\mathcal{N}}(x, y) = \min\{s(I_1(x, y), I_2(\mathcal{N}(x, y))\}. \tag{4}$$

Figure 4 shows images pairs after registration and a derived image showing the SSIM at each pixel position. The examples are the same as shown in Fig. 2.

5 Evaluation

The goal of quality assurance in digital preservation is to reduce the manual interaction and assessment. Therefore, automatic assessment using the suggested procedure is used and a small subset of image pairs with low average SSIM are interactively checked. The best average SSIM and also the smallest set size for human assessment is obtained by using all keypoints in matching. The average number of keypoints per image was 32352 for the considered data set. Sampling of robust keypoints from all available keypoints offers the lowest possibility to obtain a small average SSIM due to mismatching followed by misregistration. On the other hand, matching of keypoints is the most time consuming part of the suggested algorithm and selection of a locally distinctive set of keypoints reduces the matching effort at the cost of increasing the subset of image pairs for manual assessment.

5.1 Dataset

The dataset is a sample of 1560 image-pairs of the International Dunhang Project (IDP). The project focuses on preserving and cataloging forty thousand manuscripts, paint-

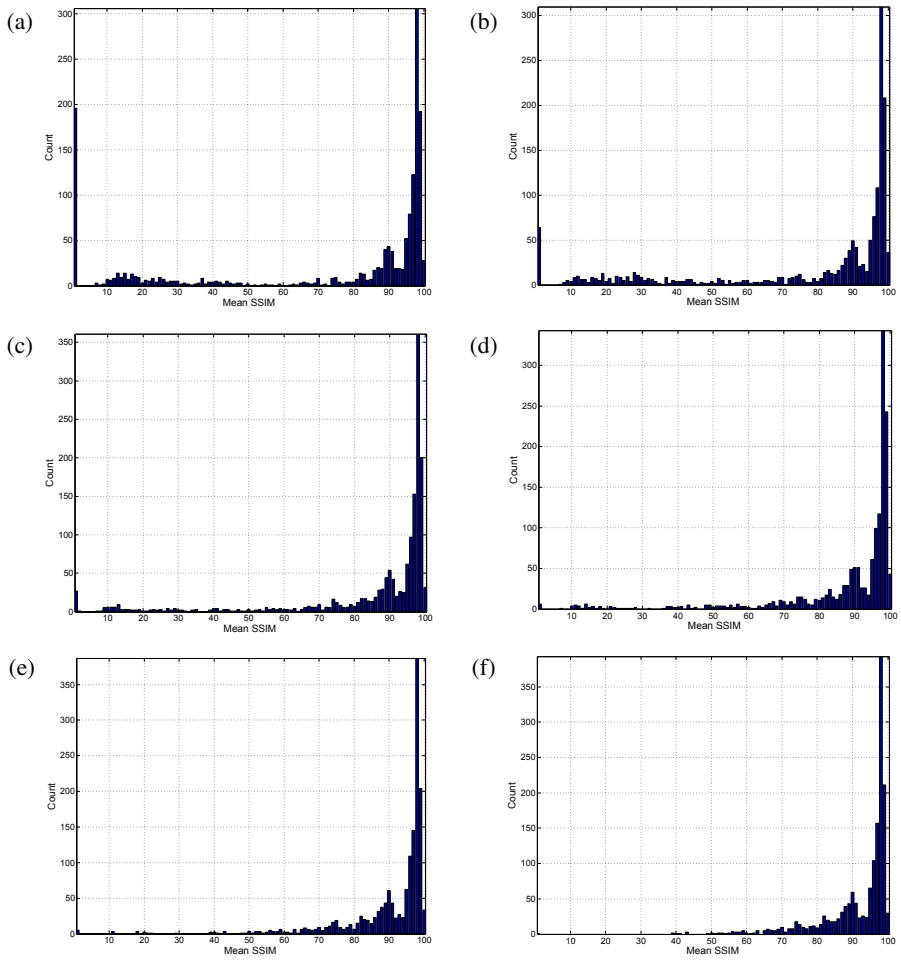

Fig. 5. Histograms of mean SSIM for different number of locally distinctive keypoints: (a) 64, (b) 256, (c) 512, (d) 1024, (e) 2048 and (f) all keypoints.

ings and printed documents dating back to the end of the first millennium. The documents were discovered in 1900 in a sealed Buddhist cave near Dunhang in western China and are now mostly dispersed to institutions worldwide. IDP started digitizing the manuscripts in 1998 to reassemble virtually the collections and make them accessible to all people.

The sample contains images of handwritten Chinese text and drawings as well as empty pages. Each image pair represents the same content, but has been digitized with different equipment and differs in size, color and alignment of the artifact.

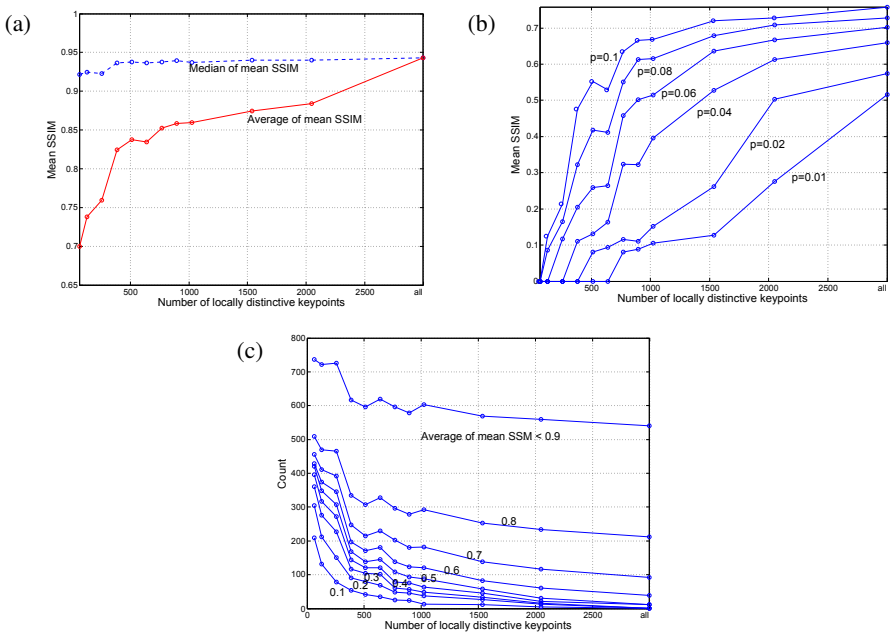

Fig. 6. Dependency of SSIM on the number of locally distinctive keypoints: (a) average and median values taken over the mean SSIM for all pairs of images, (b) quantiles on the average value of mean SSIM for all pairs of images, (c) Number of images below different thresholds on the average value of mean SSIM for all pairs of images.

6 Results

We will study the number of images with low values of mean SSIM depending on the number of keypoints. The distribution of the mean SSIM for different settings of the number of interest regions is shown in Fig. 4. Note that the number of keypoints is equal to the number of interest region provided that at least one keypoint is identified in each interest region.

For small numbers of 64 to 512 keypoints, see Figs. 4 (a)-(c), some entries for low mean SSIM are observed in the corresponding histograms. For larger numbers numbers of 1024 or 2048 keypoints, see Figs. 4 (d) and (e), the distributions already become close to the one observed for all keypoints used in Fig. 4 (f).

Figure 6 (a) shows the behavior of median and average values of mean SSIM depending on the number of keypoints. The median value of the mean SSIM is already quite high for small numbers of keypoints, e.g. for 384 keypoints. Figure 6 (b) presents the dependency of the mean SSIM with respect to p-quantiles, e.g. the p=0.1 shows the best value of the mean SSIM for the lower ranking 10 percent of the image pairs. For p-quantiles larger than 5 a reasonable good mean SSIM value of 0.6 is obtained for keypoint sets larger than 1500. Figure 6 (c) presents observed numbers of keypoints which might be left to human inspection when thresholding on the mean SSIM value

is applied. The numbers of images with low mean SSIM depending on the number of keypoints is observed from this plot, e.g. if one is interested in less than 100 images to check and a SSIM of 0.5 is assumed to be sufficiently good, at least 1024 keypoints per image were required.

7 Conclusion

We have presented an approach to image based quality assurance for digital collections. It enables automatic quality assessment of digitized documents using spatially distinctive local keypoints and robust symmetric descriptor matching. Experimental results showed, that structural similarity can be calculated from documents regardless their visual content (e.g. text, images, mixed). It is even possible to assess reliable values for images of blank old book pages. Though results showed, that the structured similarity improves on the number of keypoints, we noticed, that a high SSIM can already be achieved at a relative small number of keypoints. Through the selection of spatially distinctive local keypoints, we could reduce the number of image pairs that have to be manually checked while reducing the number of keypoints.

8 Acknowledgment

The image data was kindly provided by the British Library. Details of the International Dunhang Project can be found at http://idp.bl.uk/

References

1. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: SURF: Speeded up robust features. Computer Vision and Image Understanding (CVIU) 110(3), 346–359 (2008)
2. van Beusekom, J., Keysers, D., Shafait, F., Breuel, T.: Distance measures for layout-based document image retrieval. In: Second International Conference on Document Image Analysis for Libraries, 2006. DIAL '06. pp. 231–242 (April 2006)
3. van Beusekom, J., Shafait, F., Breuel, T.: Image-matching for revision detection in printed historical documents. In: Pattern Recognition. Proceedings of the 29th Annual Symposium of the German Association for Pattern Recognition (DAGM-2007). LNCS, vol. 4713, pp. 507–516. Springer (Sep 2007)
4. Breuel, T.: Fast recognition using adaptive subdivisions of transformation space. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92. pp. 445 –451 (Jun 1992)
5. Chaudhury, K., Jain, A., Thirthala, S., Sahasranaman, V., Saxena, S., Mahalingam, S.: Google newspaper search - image processing and analysis pipeline. In: 10th International Conference on Document Analysis and Recognition, 2009. ICDAR '09. pp. 621 –625 (July 2009)
6. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV. pp. 1–22 (2004)
7. Doermann, D., Li, H., Kia, O.: The detection of duplicates in document image databases. Image and Vision Computing 16(12-13), 907 – 920 (1998)

8. Ferrari, V., Tuytelaars, T., Gool, L.V.: Simultaneous object recognition and segmentation from single or multiple model views. Intl. J. of Comp. Vis. 67(2), 159–188 (2006)
9. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 381–395 (June 1981)
10. Gabarda, S., Cristóbal, G.: Blind image quality assessment through anisotropy. J. Opt. Soc. Am. A 24(12), B42–B51 (Dec 2007)
11. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. of ALVEY Vision Conf. pp. 147–152 (1988)
12. Ke, Y., Sukthankar, R., Huston, L.: An efficient parts-based near-duplicate and sub-image retrieval system. In: Proceedings of the 12th annual ACM international conference on Multimedia. pp. 869–876. MULTIMEDIA '04, ACM, New York, NY, USA (2004)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. of Comput. Vision 60(2), 91–110 (2004)
14. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. on Pat. Anal. and Mach. Intel. 27(10), 1615–1630 (2005)
15. Moorthy, A., Bovik, A.: Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Transactions on Image Processing 20(12), 3350 –3364 (Dec 2011)
16. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., Romeny, B.T.H., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing 39 (1987)
17. Ramachandrula, S., Joshi, G., Noushath, S., Parikh, P., Gupta, V.: Paperdiff: A script independent automatic method for finding the text differences between two document images. In: The Eighth IAPR International Workshop on Document Analysis Systems, 2008. DAS '08. pp. 585 –590 (Sep 2008)
18. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: European Conference on Computer Vision. vol. 1, pp. 430–443 (May 2006)
19. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: An efficient alternative to SIFT or SURF. In: International Conference on Computer Vision. Barcelona (Nov 2011)
20. Wang, Z., Bovik, A.: A universal image quality index. IEEE Signal Processing Letters 9(3), 81 –84 (Mar 2002)
21. Wang, Z., Bovik, A.: Mean squared error: Love it or leave it? A new look at signal fidelity measures. IEEE Signal Processing Magazine 26(1), 98 –117 (Jan 2009)
22. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600 –612 (April 2004)
23. Wu, X., Zhao, W.L., Ngo, C.W.: Near-duplicate keyframe retrieval with visual keywords and semantic context. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval. pp. 162–169. CIVR '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1282280.1282309
24. Xu, D., Cham, T.J., Yan, S., Duan, L., Chang, S.F.: Near duplicate identification with spatially aligned pyramid matching. Circuits and Systems for Video Technology, IEEE Transactions on 20(8), 1068 –1079 (Aug 2010)
25. Zhang, L., Zhang, L., Mou, X., Zhang, D.: FSIM: A feature similarity index for image quality assessment. IEEE Transactions on Image Processing 20(8), 2378 –2386 (Aug 2011)
26. Zhao, W.L., Ngo, C.W., Tan, H.K., Wu, X.: Near-duplicate keyframe identification with interest point matching and pattern learning. Multimedia, IEEE Transactions on 9(5), 1037 –1048 (Aug 2007)