An image based approach for content characterization in document collections

Reinhold Huber-Mörk and Alexander Schindler

Intelligent Vision Systems, Safety & Security Department
AIT Austrian Institute of Technology GmbH
Vienna, Austria
Email: reinhold.huber-moerk@ait.ac.at

Abstract. We consider the task of content based analysis and categorization in large-scale historical book scanning projects. Mixed content, deprecated language, noise and unexpected distortions suggest an image based approach. The use of keypoint extractors combined with the bag of features approach is applied to scanned text documents. In order to incorporate spatial information into the bag of features approach we consider three methods of spatial verification. An approach based on comparison of statistical properties of local keypoint properties such as size orientation and scale showed comparable quality in content comparison while being computationally much more efficient. Cluster analysis delivers groups of pages characterized by common properties, especially duplicated page content is detected with high reliability.

1 Introduction

Approaches for fast access to digital-born or digitized modern documents are successfully applied on the web and in modern document workflows employing information retrieval techniques. We investigate content categorization and comparison in large scale scanning projects of historical books [16] and newspapers [3]. The large amount of visual data raises issues of automatic indexing, quality assurance and information extraction based on image processing. A common approach to document image analysis is to index and compare pages based on textual information extracted through Optical Character Recognition (OCR). This method is quite limited with respect to accuracy and flexibility, especially when taking into account historical documents that are typically characterized by mixed content, deprecated language, annotations, noise and unexpected distortions. In fact, non-textual content is sometimes predominating and contains significant information (e.g. stamps, handwritten remarks etc.).

In general, several approaches for the identification of individual objects in large image collections have been proposed in the literature. Near-duplicate detection of keyframes using one-to-one matching of local descriptors was described for video data [26]. A bag of features (BoF) [5] 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 applied [14]. Doermann et. al [6] discussed the problem of duplicate detection in large document image archives and pointed out the advantage of an image recognition based approach. The approach taken at this time was to base the analysis on the shape of characters. Beusekom et. al [21] address the analysis of different versions of scanned historical documents. Baluja and Covell [1] describe an approach to differentiate between text and image content, especially line drawings, in scanned document pages.

Image based approaches can be used for detection of image content and are commonly based on local image feature descriptors. One of the most prominent local keypoint detection and description methods is the Scale Invariant Feature Transform (SIFT) [17]. SIFT operates on a scale space representation for feature detection from local gradient distributions. The SIFT descriptor is an invariant representation of local image content used in image matching, recognition or comparison frameworks. The BoF derived from local descriptors such as SIFT was described as an efficient approach to content based retrieval and detection from image data. The BoF approach is inspired by the bag of words approach based on term frequency weighting and comparison in text retrieval [18]. SIFT features are combined with AdaBoost learning to obtain relevant information in large-scale book-scanning systems, e.g. preview pages. Discrimination of main body textand decorative elements in historical manuscripts using SIFT and a support vector machine (SVM) classifier is described by Garz et. al [9].

The most popular approach when comparing image content based on the BoF approach is to use term frequencies (tf) and inverse document frequencies (idf) of visual words. Jégou et. al pointed out weaknesses of the idf scheme related to burstiness of visual features, i.e. multiple occurrence of a specific visual words in one image or specific visual words occurring among many images [13]. For text we observed burstiness of visual words occurring at similar spacings between characters at fine scales and for pieces of characters at very fine scales. Combination with a higher level description, e.g. using visual word co-occurrence matrices [25] or co-ocsets [4], by spatial subdivision such as spatial pyramid matching [24], or verification based on matching in the image space [15] try to overcome limitations of a purely tf based approach.

Detection of duplicated pages in an automatic document scanning workflow was investigated for historical documents where a BoF based on SIFT descriptors approach was chosen and results were compared to manually obtained data [12] This paper demonstrates image content grouping by similarity using a BoF approach combined with spatial verification schemes. Groups of similar pages, e.g. layout or graphical properties etc., could be identified as well as duplicated content.

The remainder of this paper is organized as follows. Section 2 describes methods for visual content comparison and introduces our approach. Results are presented in Section 3 and Section 4 summarizes the paper.

2 Image comparison

To detect and describe interest regions in document images we used the SIFT keypoint extraction and description approach. Subpixel image location, scale and orientation are associated with each SIFT keypoint. The associated SIFT descriptor consists of a 4×4 location grid containing 8 gradient orientation bins in each grid cell. The descriptor vectors of length 128 will be used to learn a visual dictionary, i.e. the BoF. Spatial verification becomes important as the BoF does not represent spatial relationships between the visual words present in an image. We will compare three methods for spatial verification: (1) estimation of a homography and comparison in the image domain, (2) comparison of the co-occurrence statistics of the visual words for two images and (3) global detector statistics comparison.

Learning of the visual dictionary is performed using a clustering method applied to all SIFT descriptors of all images, which could become computationally very demanding. As a single scanned book page already contains a large number of local descriptors we applied preclustering of descriptors to each image. In contrast to a similar procedure, where all descriptors for all images of the same category are clustered independently and subsequently appended to the BoF [10], we construct a list of clustered descriptors for each page and cluster this list in a second step in order to obtain a dictionary for the whole book. We used k-means for preclustering and final clustering of the BoF. Individual terms i occur on each page with varying frequency t_i. The visual histogram of term frequencies t_i for an individual book page is derived from the BoF representation by counting the indices of the closest descriptors. The term frequencies t_i are represented in its normalized form, i.e. $\sum_{i=1\ldots|V|} t_i = 1$, where V is the set of visual words contained in the visual vocabulary for an individual book. Matching of two visual term frequency histograms t^a and t^b is based on histogram intersection $T_{ab} \in [0, 1]$ given by

$$T_{ab} = \sum_{i=1}^{|V|} \min(t_i^a, t_i^b). \tag{1}$$

To group image with respect to similarity we first calculate the similarity measure T_{ab} for each page a to all other pages b in the collection B. Taking the maximum of $T_{ab}, \forall b \in B, b \neq a$ delivers a view of collection consistency, i.e. if all T_{ab} are similar the document content is quite homogeneous and if T_{ab} shows different modes the content and page structure is supposed to be mixed. Figure 1(a) shows a plot for T_{ab} for an example book-scan consisting of 256 pages. The main body of the book receives a maximum T_{ab} of around 0.6 which basically is related the self-similarity of the text pages. Bursts exceeding this value are typically duplicated pages. Single maximum peaks correspond to very similar pages of low noise, e.g. empty pages. Peaks of low similarity measure are pages not similar to any other content, e.g. these are often the cover pages.

Spatial verification is applied to each page a using a shortlist delivered by ranking the similarity T_{ab}. Details of combination of term frequency matching

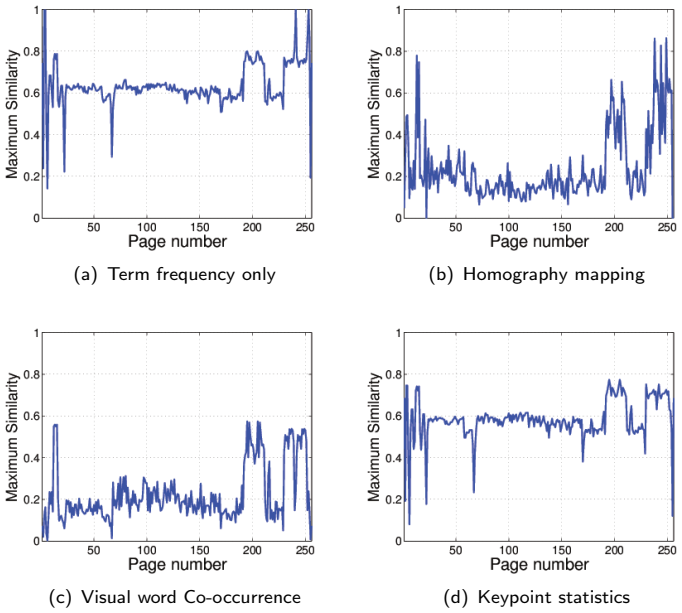


Fig. 1. Maximum similarity for each image of a book collection.

with three different approaches for spatial verification is described in the following.

2.1 Homography estimation and image similarity

An affine transformation was found sufficient to overlay images obtained by current book-scan devices as the main problem of bent pages typically occurring with flatbed scanners is not observed. Matching of SIFT features uses the RANSAC procedure for robust estimation [8]. In order to limit the complexity of the matching procedure spatial subsampling of keypoints by identifying the most salient keypoints with respect to a spatial grid was employed [11]. The similarity of two overlaid images is expressed by the mean structural similarity index (SSIM) [22], where a mean SSIM $\rightarrow 1$ indicates identical content and an SSIM $\rightarrow 0$ means unrelated content. Figure 1(b) shows the similarity plot for the considered example.

2.2 Co-occurrence of visual words

We follow the basic ideas concerning descriptive visual words as described by Zhang et. al [25]. For each image co-occurrence matrices $c(v_s, v_t), \ldots v_s, v_t \in V$ counting the concurrent appearance of visual words v_s and v_t in a spatial neighborhood are constructed. Each keypoint is assigned to a visual word and

delivers a contribution to the co-occurrence statistics by counting the concurrent presence of the visual words assigned to of all keypoints contained in its spatial neighborhood. The spatial neighborhood is a circular region around a keypoint location depending on the corresponding keypoint size delivered by SIFT. The size d of the influence region for a keypoint was chosen as $d = s \cdot p_d$, where s is the estimated size of the SIFT keypoint and p_d the scaling parameter. A selection of $p_d = 6$ is derived from the the spatial support of the SIFT detector. Visual word co-occurrence matrices c^a and c^b are normalized $\sum_{s...|V|} \sum_{t=1...|V|} c(v_s, v_t) = 1$ and matched using 2D-histogram intersection $C_{ab} \in [0, 1]$

$$C_{ab} = \sum_{s=1}^{|V|} \sum_{t=1}^{|V|} \min(c^a(v_s, v_t), c^b(v_s, v_t)). \tag{2}$$

The most crucial part of this approach is to identify spatially adjacent neighboring keypoints for each keypoint. The k-d tree representation allows efficient representation and queries of spatially local neighborhoods [2]. Nevertheless, the computational demands of this method were found on the order of magnitude when compared to homography based image comparison. Figure 1(c) shows the similarity plot for the considered example.

2.3 Global keypoint property statistics

We suggest a method based on global statistics of keypoint properties. SIFT delivers location, size and orientation for each keypoint. In the following we describe how to use these properties for spatial verification.

We use a measure of inhomogeneity to characterize the spatial distribution of keypoints [20]. The image is subdivided into a sequence $s^2, s = 1, 2, 4, \ldots$ of rectangular regions of equal size and the number of keypoints m_i falling into region $i, i = 1, \ldots, s^2$ is obtained

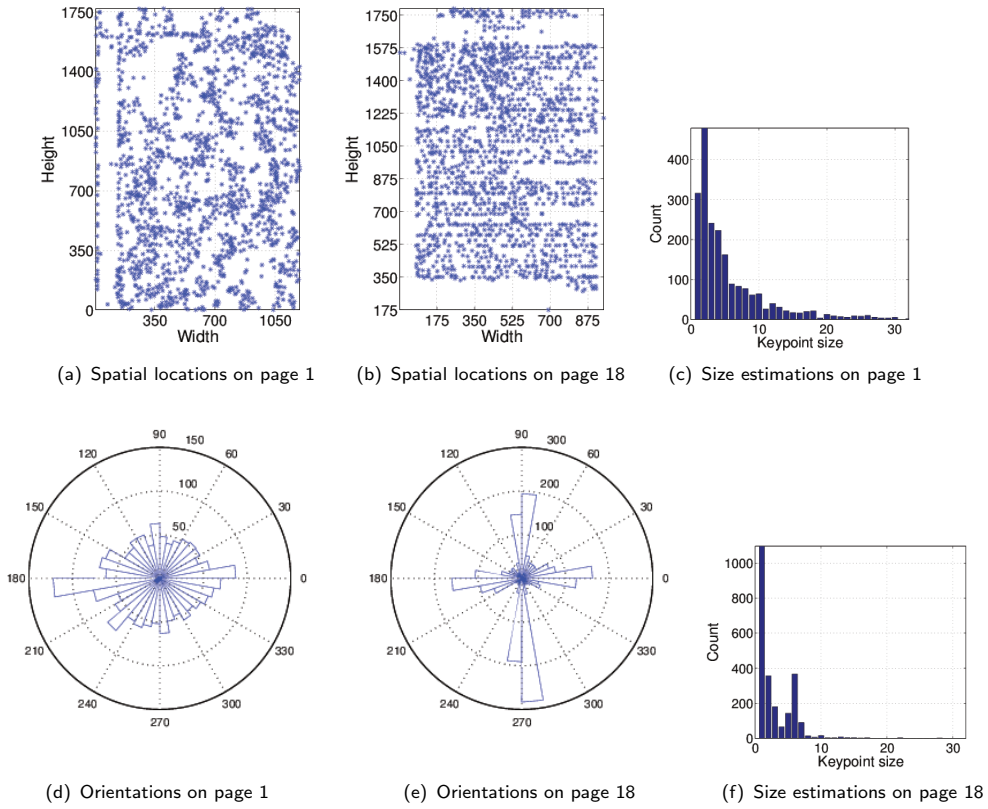
$$h = \sum_{j=1}^{\log_2 s} w^{1-j} h(2^j), \quad h(s) = \frac{1}{2n} \sum_{i=1}^{n^2} |m_i - \frac{n}{s^2}|, \tag{3}$$

where $w = 4.79129$ was derived in [20]. Images having spatially uniformly distributed keypoints obtain values of $h \to 0$ and whereas for spatially concentrated keypoints we get $h \to 1$.

We exploit the orientation estimation delivered by SIFT using a measure for circular uniformity U. The U measure was introduced by Rao [19] in his test for circular uniformity

$$U = \frac{1}{2}\big(\sum_{i=1}^{n-1} |(\alpha_{i+1} - \alpha_i) - \lambda| + |(360 - \alpha_n + \alpha_1) - \lambda|\big). \tag{4}$$

where $\lambda = 360/n$ and the angle directions $0 \leq \alpha_i < 360$ are sorted in ascending order $\alpha_{i+1} > \alpha_i, \forall i = 1 \ldots, n-1$. We normalize the uniformity measure to the



(a) Spatial locations on page 1 (b) Spatial locations on page 18 (c) Size estimations on page 1

(d) Orientations on page 1 (e) Orientations on page 18 (f) Size estimations on page 18

Fig. 2. Statistical characterization of keypoints derived from spatial, circular and scale properties for two different book pages.

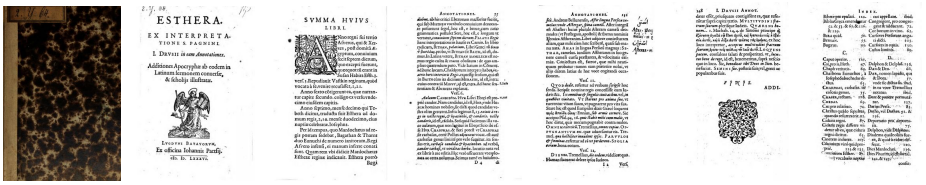
range $[0, 1]$ by $u = U/360$. Images with keypoints directions pointing uniformly into all directions obtain $u = 0$ and coherently oriented keypoints we get $u \to 1$.

We have chosen the variance of the keypoint size S as a descriptor for the distribution of the size estimations over all detected keypoints. A normalized version s is obtained from the maximum size of a SIFT keypoint

$$s = S/(\sigma_0 2^{o_{\max}+(s_{\max}-1)/s_{\max}}), \tag{5}$$

where σ_0 is the initial Gaussian smoothing parameter. The maximum scale index is denoted by $s_{\max}$ and is typically set to 4. The maximum octave index is given by $o_{\max}$ and depends on the image resolution.

Figure 2 shows spatial radial and size distributions for keypoint properties. The plots on the left correspond to the image shown in Figure 3(a) and $h_1 = 0.13335$, $U_1 = 226.2887$ and $S_1 = 3.8964$ were calculated. The plots on the right correspond to the image shown in Figure 3(c) and $h_{18} = 0.2116$, $U_{18} =$



(a) Page 1 (b) Page 5 (c) Page 18 (d) Page 77 (e) Page 153 (f) Page 170 (g) Page 232

Fig. 3. Scanned sample pages of a historical book.

274.6681 and $S_{18} = 2.2229$ were calculated, i.e. text content tends to be spatially homogeneous, circular less uniform and has lower scale variation.

2.4 Combination of term frequency with spatial verification

Term frequency based comparison was used to deliver a shortlist L, we used a size of $|L| = 3$ in our experiments. Based on this shortlist we applied spatial verification and combined term frequency matching and spatial matching in a conjunctive fashion, i.e. the final similarity measure is derived from $T_{ab} \cdot V_{ab}, b \in L$ where V_{ab} is either $V_{ab} = SSIM$ in the case of homography based verification, $V_{ab} = C_{ab}$ in the case of co-occurrence based verification and the following heuristics $V_{ab} = 1 - \sqrt{(|h_a - h_b| + |u_a - u_b| + |s_a - s_b|)/3}$ is used for the approach based on keypoint property statistics.

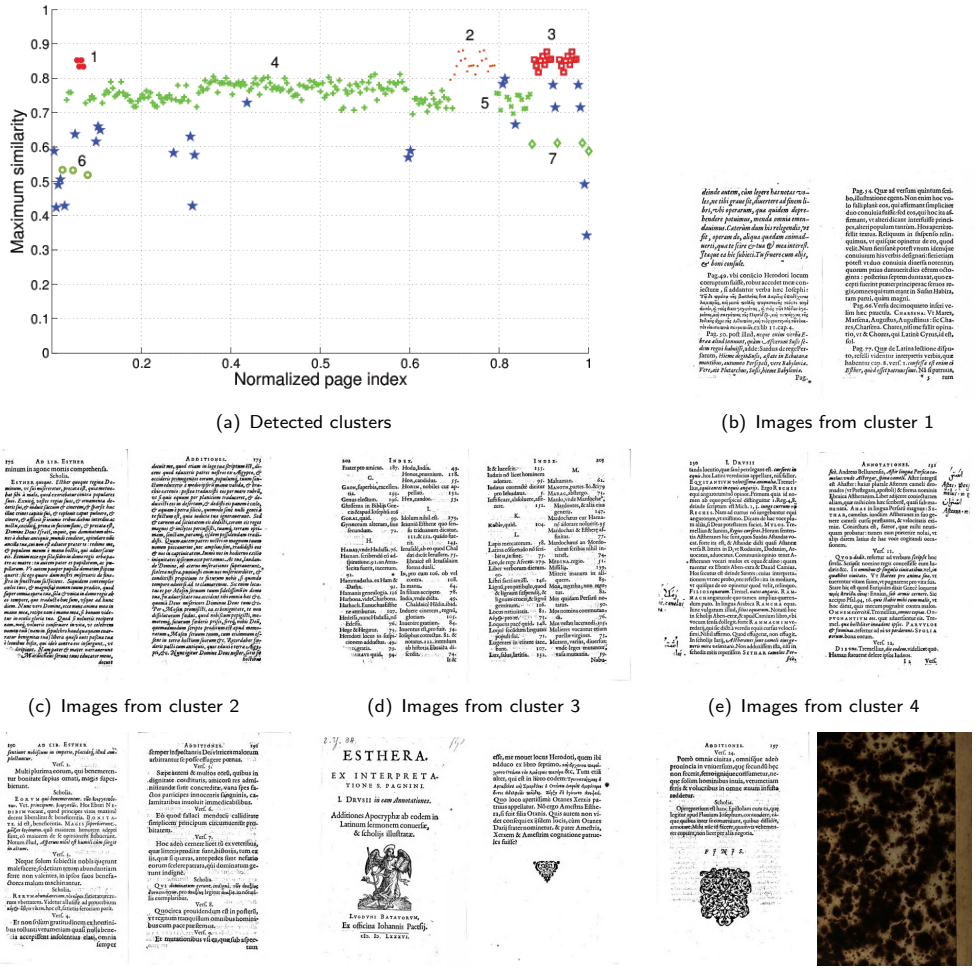
3 Results

We consider collections of historical books obtained by an automated book-scanning device, see Figure 3 for sample pages extracted from an exemplary book. A BoF for each scanned book is constructed and visual term histograms for each page are extracted. A combination with spatial verification is performed as described in Sec. 2.4. It was observed, that page content and structure also depends on the absolute page count. Furthermore, when considering the case of duplicated pages it was observed that duplicated pages occur blockwise. This occurs due the operations of the automatic book scan device, which turns back a number of pages in situations of a possible error.

We performed an unsupervised clustering of the space spanned by maximum similarity for each page and the normalized page count. Identified clusters were characterized by similar page content or layout. We used the DBSCAN algorithm [7] to discover clusters in the page similarity/index space. Figure 4(a) shows the result of DBSCAN, where seven clusters have been identified. The isolated star-shaped points are outliers and represent unique content. The main body of the book is covered by clusters 4 and 5, example pages are shown in Figure 4(d) and Figure 4(e). Duplicated pages are detected from clusters 1, 2 and 3, e.g. Figure 4(b) shows two scanned pages of same content with small differences

in skew and noise. Duplicate detection delivered correct results when visually verified. The regions of duplications are automatically obtained by thresholding of the similarity measure [12].

Run time measurements were performed on a Xeon 3.6 GHz computer using a MATLAB/C implementation. To analyze a scanned book with 256 pages scanned at 72 DPI (1800 × 1200 pixel) it took 31 seconds for tf matching only, combined matching based on homography estimation and image comparison took 449 seconds, combination with co-occurrence verification took 451 seconds and combined matching using keypoint property statistics consumed 128 seconds.



(a) Detected clusters (b) Images from cluster 1

(c) Images from cluster 2 (d) Images from cluster 3 (e) Images from cluster 4

(f) Images from cluster 5 (g) Images from cluster 6 (h) Images from cluster 7

Fig. 4. Clustering of page similarity/index plane and sample images from different clusters.

4 Conclusion

We have presented an approach for visual page content clustering applicable to duplicate detection in book-scan systems. A BoF approach combined with spatial verification based on keypoint statistics was found suited for the analysis of scanned historical book collections. Background knowledge on the sequential nature of the scanning process and incorporation of spatial knowledge using global keypoint statistics improves results significantly. Clusters of duplicated content are automatically detected and subject manual quality assurance in a library workflow. The system is currently evaluated for the task of content characterization and duplicate detection at the Austrian National Library. Future research includes content classification with respect to image quality categories.

Acknowledgment

The authors would like to thank Sven Schlarb from the Austrian National Library (ONB) for providing data and expertise on library workflows.

This work is part of the SCAlable Preservation Environments (SCAPE) project which aims at developing scalable services for planning and execution of institutional digital preservation strategies. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement no. 270137).

References

1. Shumeet Baluja and Michele Covell. Finding images and line drawings in document-scanning systems. In *Proc. Intl. Conf. on Document Analysis and Retrieval ICDAR'09*.
2. Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, Sep 1975.
3. K. Chaudhury, A. Jain, S. Thirthala, V. Sahasranaman, S. Saxena, and S. Mahalingam. In *Proc. Intl. Conf. on Document Analysis and Recognition ICDAR'09*.
4. Ondrej Chum and Jiri Matas. Unsupervised discovery of co-occurrence in sparse high dimensional data. In *Proc. Computer Vision and Pattern Recogntition CVPR'10*.
5. Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and C?dric Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV'04*.
6. David Doermann, Huiping Li, and Omid Kia. The detection of duplicates in document image databases. *Image and Vision Computing*, 16(12-13):907–920, 1998.
7. Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. Conf. on Knowledge Discovery and Data Mining KDD'96*.
8. Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, Jun 1981.

9. Angelika Garz, Robert Sablatnig, and Markus Diem. Layout analysis for historic manuscripts using SIFT features. In *Proc. Intl. Conf. on Document Analysis and Recognition ICDAR'11*.
10. Lykele Hazelhoff, Ivo Creusen, Dennis van de Wouw, and Peter H. N. de With. Large-scale classification of traffic signs under real-world conditions. In *Proc. SPIE Electronic Imaging: Algorithms and Systems VI*, 2012.
11. Reinhold Huber-Mörk and Alexander Schindler. Quality assurance for document image collections in digital preservation. In *Proc. Advanced Concepts for Intell. Vision Sys. ACIVS'12*, volume 7517 of *Springer LNCS*, 2012.
12. Reinhold Huber-Mörk, Alexander Schindler, and Sven Schlarb. Duplicate detection for quality assurance of document image collections. In *Proc. Conf. on Digital Preservation iPres'12*.
13. Hervé Jégou, Matthijs Douze, and Cordelia Schmid. On the burstiness of visual elements. In *Proc. Computer Vision and Pattern Recognition CVPR'09*.
14. Yan Ke, Rahul Sukthankar, and Larry Huston. An efficient parts-based near-duplicate and sub-image retrieval system. In *Proc. Intl. Conf. on Multimedia MULTIMEDIA'04*.
15. J. Knopp, J. Sivic, and T. Pajdla. Avoiding confusing features in place recognition. In *Proc. Europ. Conf. on Computer Vision ECCV'10*.
16. Adam Langley and Dan S. Bloomberg. Google books: making the public domain universally accessible. In *Proc. of SPIE, Document Recognition and Retrieval XIV*, 2007.
17. David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. of Comput. Vision*, 60(2):91–110, 2004.
18. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 7 edition, 2008.
19. J. S. Rao. Bahadur efficiencies of some tests for uniformity on the circle. *Ann. Math. Statist.*, 43(2):468–479, 1972.
20. U. Schilcher, M. Gyarmati, C. Bettstetter, Yun Won Chung, and Young Han Kim. Measuring inhomogeneity in spatial distributions. In *Proc. Vehicular Technology Conference, VTC'08*.
21. Joost van Beusekom, Faisal Shafait, and Thomas Breuel. Image-matching for revision detection in printed historical documents. In *Proc. Symp. of German Assoc. for Pattern Recognition*, volume 4713 of *LNCS*, 2007.
22. Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Proc.*, 13(4):600–612, Apr 2004.
23. Xiao Wu, Wan-Lei Zhao, and Chong-Wah Ngo. Near-duplicate keyframe retrieval with visual keywords and semantic context. In *Proc. Conf. on Image and Video Retrieval CIVR'07*.
24. Dong Xu, Tat Jen Cham, Shuicheng Yan, Lixin Duan, and Shih-Fu Chang. Near duplicate identification with spatially aligned pyramid matching. *IEEE Trans. Circuits Syst. Video Techn.*, 20(8):1068 –1079, Aug 2010.
25. Shiliang Zhang, Qi Tian, Gang Hua, Qingming Huang, and Shipeng Li. Descriptive visual words and visual phrases for image applications. In *Proc. Intl. Conf. on Multimedia MULTIMEDIA'09*.
26. Wan-Lei Zhao, Chong-Wah Ngo, Hung-Khoon Tan, and Xiao Wu. Near-duplicate keyframe identification with interest point matching and pattern learning. *IEEE Trans. Pat. Anal. Mach. Intell.*, 9(5):1037 –1048, Aug 2007.