Retrieving Writing Patterns From Historical Manuscripts Using Local Descriptors

Rainer Herzog, Arved Solth, Oliver Bestmann, Julian Scheel, Bernd Neumann

1. Introduction

Computer-supported retrieval of manuscripts based on the visual features of a query image is a highly desirable, but rarely available service for manuscript research. The service could be used, for example, to check whether a manuscript, specified by a copy, is contained in a museum collection. This kind of retrieval is often approximated by making use of an index based on textual annotations, and thus requires extensive manual preparation. Retrieval based on a query image without annotations, on the other hand, promises to be mainly automatic and also support interesting applications beyond document retrieval. Most importantly, this service can allow retrieval of manuscripts containing the query image as a detail. For example, one could find manuscripts where characters are written in a specific way, exemplified in the query image. Moreover, one could search for the occurrence of writing patterns consisting of arbitrary graphical features, in short graphs. Similar graphs, retrieved from different manuscripts, may contribute valuable information about a possible scribe identity or a common origin of manuscripts.

In this contribution[1] we describe a novel approach for graph retrieval based on local descriptors at "interest points". Interest points (IPs) specify locations of strong "cornerness" of the image intensities and thus provide reasonably stable reference points for local descriptions. They have proved their worth in many image analysis applications, in particular in image retrieval solutions based on SIFT features [1]. Different from SIFT features, our descriptors are not scale and rotation invariant, although tolerant to small variations, giving rise to a distinctly superior performance and efficiency while preserving its usefulness for many concerns of manuscript research. Basically, each descriptor consists of the structure tensors [2] in the neighborhood of an IP, thus giving a precise account of the local gradient distribution. Depending on the resolution of the manuscripts and the chosen size of the neighborhood, the descriptor of a single IP may comprise several hundred feature values, called *IP features* in the sequel. For highly detailed query images, for example depicting a Chinese character, there may be a large number of IPs (in our experiments up to 40), each of which is recorded with its location and its feature values in a local descriptor.

For retrieval, a target image is processed essentially in the same way as the query image, i.e. IPs and IP features are determined. This is done only once and off-line, comparable to establishing an annotation. The main retrieval task is then to find a subset of IPs in a target image whose relative locations and feature vectors best match the query descriptors. We will present an approach which profits from prior segmentation of the target image into possible matching candidates, but can also be applied to large datasets without segmentations. It is controlled by a simple probabilistic model for the kind of differences between query and data which should be tolerated for a retrieval, with parameters which can be adjusted by a manuscript researcher.

---

[1] This work was supported by the DFG Research Group "Manuscript Cultures in Asia and Africa" and the DFG Collaborative Research Center for the Study of Manuscript Cultures SFB 950.

First experimental results with Chinese characters in historical manuscripts indicate that our descriptor is tolerant with respect to a certain amount of variations of the same character, yet quite discriminative with respect to structurally different characters, promising high precision and recall.

In the remainder of this abstract, we describe related work in Section 2, give details about the technical implementation in Section 3, and finally report about experimental results in Section 4.

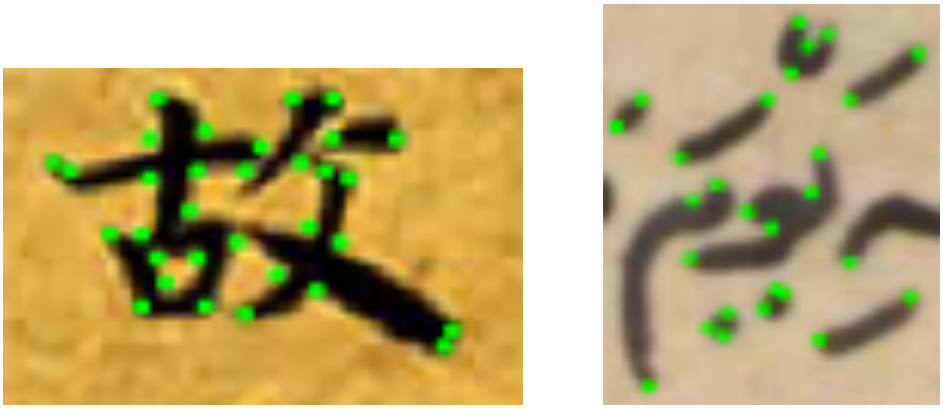
2. Related Work

Retrieving graphs from manuscripts is a special case of Content-Based Image Retrieval (CBIR), a well-established research field with a rich set of methods [3]. But CBIR applied to manuscripts has found very little attention in this community. Most relevant work builds on handwriting recognition which is increasingly applied to historical manuscripts [4, 5, 6, 7]. Most approaches so far use retrieval based on words or characters [6, 7, 8, 9] which limits the applicability to handwritings with clear word or character separation. A more general approach, as followed in this contribution, relies on retrieving a spatial configuration of graphical features, extracted from the query [10]. For all approaches, the key question is which features to use for the comparison of query and data. It has been shown that the spatial distribution of gradients gives the best results [9, 11]. In most approaches, descriptors based on simple gradient computations [8, 11] are assigned to a fixed mesh covering a segment. In our work, descriptors are based on structure tensors [2] at IPs determined for each query independent of a segmentation. Furthermore, we use a novel probabilistic model applicable to IPs.

3. Technical Implementation

Due to the restricted length of this abstract, we cannot provide many technical details here. The IPs in our approach are computed using the Harris Corner Detector [12] with the improvements introduced in [2]. Fig. 1 shows typical results achieved for manuscripts with Chinese and Arabic handwritings.

Local descriptors are computed for each IP by combining the structure tensors of all points in the neighbourhood (in our examples of size 11 x 11) into a large feature vector characterizing strength and directionality of intensity gradients.



a b

Fig. 1: Interest points (IPs) determined (a) for a Chinese character in a 90 x 50 image and (b) for a 50 x 60 segment of Arabic handwriting

To retrieve matching patterns from target data, IPs of the query are incrementally compared with IPs of the data using a probabilistic model comprising the following boolean probabilities, both for the hypothesis H_1 that the target is a match, and the hypothesis H_0 that it is not a match:

P_{IP} target descriptor missing in window of query location

P_{loc} target descriptor dislocated from query location

P_{feat} target descriptor features differing from query descriptor features

Let A be pairs of descriptors of query and target, $P(H_0)$ and $P(H_1)$ the prior probabilities of the respective hypotheses, then for a hypothesis test we have to evaluate:

$$\frac{P(A|H_1)}{P(A|H_0)} = \frac{P_{IP}(A|H_1)\,P_{loc}(A|H_1)\,P_{feat}(A|H_1)}{P_{IP}(A|H_0)\,P_{loc}(A|H_0)\,P_{feat}(A|H_0)} > \frac{P(H_0)}{P(H_1)}$$

The comparison is formulated as two hypothesis tests, the first whether the query pattern is contained in the target, and the second whether the target pattern, constrained by a successful first test, is contained in the query. The target is considered a match of the query, if both tests succeed.

4. Experimental Results

We have carried out first retrieval experiments with Chinese and Arabic manuscripts. Fig. 2 left shows a section of the Fo shuo Tiwei jing (British Library Or.8210/S. 2051). The left-most green box marks a character used as a query, the other green boxes show matching characters found by the retrieval system. There have been no false negatives, but the second hit in Column 6 is a false positive, although quite similar to the query. Similar results have been achieved for other queries, applied to a manuscript section with 364 characters.

Fig. 2: Retrieval from a Chinese (left) and an Arabic manuscript (right), see text.

Fig. 2 right shows part of Vollers0015,S.2 from the Refaiya Library in Leipzig. The section marked with a box (shown also in Fig. 1b) was used as a query for the same manuscript and retrieved as the only hit as to be expected. Further experiments are on the way.

References

[1] D.G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, Int. J. of Computer Vision, 60.2, 2004, 91-110.

[2] U. Koethe: Edge and Junction Detection with an Improved Structure Tensor. Proc. 25th DAGM Symposium, 2003, 25-32.

[3] R. Datta, D. Joshi, J. Li, J.Z. Wang: Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys, Vol. 40, No. 2, Art. 5, 2008, 1-60.

[4] A. Fischer, M. Wüthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, M. Stolz: Automatic Transcription of Handwritten Medieval Documents. 15th Int. Conf. on Virtual Systems and Multimedia, 2009, 137-142.

[5] I.B. Yosef, I. Beckman, K. Kedem, I. Dinstein: Binarization, Character Extraction, and Writer Identification of Historical Hebrew Calligraphy Documents. Int. J. on Document Analysis and Recognition (IJDAR) Vol. 9, 2007, 89-99

[6] V. Lavrenko, T. Rath, R. Manmatha: Holistic Word Recognition for Handwritten Historical Documents. Proc. Document Image Analysis for Libraries (DIAL), 2004, 278-287.

[7] T. Adamek, N.E. O'Connor, N. Murphy, A.F. Smeaton: Word Matching Using Single Closed Contours for Indexing Handwritten Historical Documents. Int. J. on Document Analysis and Recognition, 9 (2-4), 2007, 153-165.

[8] S. Srihari, C. Huang, H. Srinivasan: A Search Engine for Handwritten Documents. Proc. SPIE-IS&T Electronic Imaging, 2005, 66-75.

[9] B. Zhang, S.N. Srihari, C. Huang: Word Image Retrieval Using Binary Features. Proc. Document Recognition and Retrieval XI, 2004, 45-53.

[10] T.-S. Su, T.-W. Zhang, D-J. Guan, H.J. Huang: Off-line Recognition of Realistic Chinese Handwriting Using Segmentation-free Strategy. Pattern Recognition 42.1, 2009, 167-182

[11] K. Ding, Z. Liu, L. Jin, X. Zhu: A Comparative Study of Gabor Feature and Gradient Feature for Handwritten Chinese Character Recognirion. Proc. Int. Conf. Wavelet Analysis and Pattern Recognition, 2007, 1182-1186.

[12] C. Harris, M. Stephens : A Combined Corner and Edge Detector. Proc. 4th Alvey Vision Conference, 1988, 147–151.