Automatically Identifying Join Candidates in the Cairo Genizah

Lior Wolf, Rotem Littman, Naama Mayer, Nachum Dershowitz
The Blavatnik School of Computer Science
Tel Aviv University

Roni Shweka, Yaacov Choueka
The Friedberg Genizah Project
Jerusalem, Israel

Abstract

A join is a set of manuscript-fragments that are known to originate from the same original work. The Cairo Genizah is a collection containing approximately 250,000 fragments of mainly Jewish texts discovered in the late 19th century. The fragments are today spread out in libraries and private collections worldwide, and there is an onging effort to document and catalogue all extant fragments.

The task of finding joins is currently conducted manually by experts, and presumably only a small fraction of the existing joins have been discovered. In this work, we study the problem of automatically finding candidate joins, so as to streamline the task. The proposed method is based on a combination of local descriptors and learning techniques.

To evaluate the performance of various join-finding methods, without relying on the availability of human experts, we construct a benchmark dataset that is modeled on the Labeled Faces in the Wild benchmark for face recognition. Using this benchmark, we evaluate several alternative image representations and learning techniques. Finally, a set of newly-discovered join-candidates have been identified using our method and validated by a human expert.

1. Introduction

Written text is one of the best sources for understanding historical life. The Cairo Genizah is a unique source of preserved middle-eastern texts, collected between the 11th and the 19th centuries. These texts are a mix of religious Jewish manuscripts with a smaller proportion of secular texts. To make the study of the Genizah more efficient, there is an acute demand to group the fragments and reconstruct the original manuscripts. Throughout the years, scholars have devoted a great deal of time to manually identify such groups, referred to as *joins*, often visiting numerous libraries.

Manual classification is currently the gold-standard for finding joins. However, it is not scalable and cannot be applied to the entire corpus. We suggest automatically identifying candidate joins to be verified by human experts. To this end, we employ modern image-recognition tools such as local descriptors, bag-of-features representations and discriminative metric learning techniques. These techniques are modified for the problem at hand by employing suitable preprocessing and by employing task-specific key-point selection techniques. Where appropriate, we use suitable generic methods.

(a) (b)

Figure 1. Example of a document from the Cairo Genizah. (a) The original fragment. (b) After the binarization process.

We validate our methods in two ways. The first is to construct a benchmark for the evaluation of algorithms that are able to compare the images of two leaves. Algorithms are evaluated based on their ability to determine whether two leaves are a join or not. In addition, we create a short list of newly discovered join-candidates that are the most likely, according to our algorithm's metric, and send it to a human expert for validation.

The main contributions of this work are as follows:

1. The design of an algorithmic framework for finding join-candidates. The algorithms are based on the application of local descriptors and machine learning techniques. The framework provides a high-throughput method for join finding in which human expertise is

utilized efficiently.

2. The study of suitable algorithmic details for obtaining high levels of performance for finding candidate joins. In particular, by carefully constructing our recognition method, we obtain an increase in recognition rate, at very low false-positive rates, of up to ten-fold.
3. Provide a benchmark for the evaluation of join-finding algorithms. Such a benchmark is important for evaluating such algorithms in the absence of accessible human experts.
4. The actual identification of new, unknown, joins in the Genizah corpus.

This contribution is presented in the IEEE Workshop on eHeritage and Digital Art Preservation following the Twelfth IEEE International Conference on Computer Vision (ICCV 2009). The full text can be found at the authors' website http://www.cs.tau.ac.il/~wolf/papers/genizah.pdf.

References

[1] K. Baird. Anatomy of a versatile page reader. *Proceedings of the IEEE*, 80(7), 1992.

[2] A. Bensefia, T. Paquet, and L. Heutte. Information retrieval based writer identification. In *Int. Conf. on Document Analysis and Recognition*, 2003.

[3] S. Bres, V. Eglin, and C. V. Auger. Evaluation of Handwriting Similarities Using Hermite Transform. In *Frontiers in Handwriting Recognition*, 2006.

[4] M. Bulacu and L. Schomaker. Automatic handwriting identification on medieval documents. In *Int. Conf. on Image Analysis and Processing*, 2007.

[5] R. G. Casey and E. Lecolinet. A survey of methods and strategies in character segmentation. *PAMI*, 18, 1996.

[6] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001.

[7] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[8] I. Dinstein and Y. Shapira. Ancient hebraic handwriting identification with run-length histograms. *Systems, Man and Cybernetics, IEEE Transactions on*, 1982.

[9] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6), 1981.

[10] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. UMASS, TR 07-49, 2007.

[11] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. *CVPR*, 2004.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[13] G. Leedham, S. Varma, A. Patankar, and V. Govindarayu. Separating text and background in degraded document images. In *Frontiers in Handwriting Recognition*, 2002.

[14] H. G. Lerner and S. Jerchower. The Penn/Cambridge Genizah fragment project: Issues in description, access, and reunification. *Cataloging & Classification Quarterly*, 42(1), 2006.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60, 2004.

[16] M. Panagopoulos, C. Papaodysseus, P. Rousopoulos, D. Dafi, and S. Tracy. Automatic writer identification of ancient greek inscriptions. *PAMI*, 31(8), 2009.

[17] S. C. Reif. *A Jewish Archive from Old Cairo: The history of Cambridge Universitys Genizah Collection*. Curzon Press, Richmond, England, 2000.

[18] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *CVPR*, 2005.

[19] S. N. Srihari and V. Govindaraju. Analysis of textual images using the hough transform, 1989.

[20] Y. Taigman, L. Wolf, and T. Hassner. Multiple one-shots for utilizing class label information. In *The British Machine Vision Conference*, 2009.

[21] L. Wolf, S. Bileschi, and E. Meyers. Perception strategies in hierarchical vision systems. In *CVPR*, 2006.

[22] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Faces in Real-Life Images Workshop in ECCV*, 2008.

[23] L. Wolf, T. Hassner, and Y. Taigman. The one-shot similarity kernel. In *ICCV*, 2009.

[24] D. H. Wolpert. Stacked generalization. *Neural Netw.*, 5(2), 1992.

[25] J. Zhang, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *IJCV*, 2007.