Proceedings of the International Conference of Machine Vision and Machine Learning
Prague, Czech Republic, August 14-15, 2014
Paper No. 133

Pivot-based Search for Word Spotting in Archive Documents

László Czúni, Péter József Kiss
University of Pannonia, Department of Electrical Engineering and Information Systems
Egyetem u. 10., H-8200 Veszprém, Hungary
czuni@almos.vein.hu

Ágnes Lipovits, Mónika Gál
University of Pannonia, Department of Mathematics
Egyetem u. 10., H-8200 Veszprém, Hungary
lipovitsa@almos.vein.hu

Abstract- Offline handwriting recognition of archive documents is an unsolved problem in image analysis. There are several approaches but the possible distortions of word shapes and the different noises result in relatively low recognition rates. In our paper we deal with a local feature based word spotting approach, which showed reasonable retrieval performance in previous studies with a drawback of slow running time. With the proposed approach we try to achieve similar recognition performance but with affordable running time. In our paper we propose a pivot based search algorithm where pivots are reference objects and are used for feature generation. The retrieval results of the new algorithm are very close to those of the full linear search but at about 2-3.5 times faster running times.

Keywords: OCR, Offline handwriting recognition, Scale Invariant Feature Transform - SIFT, Non-metric space, Pivot based search.

1. Introduction

There are several types of handwriting analysis, such as signature recognition, handwriting recognition, and writer identification. While online methods, using time series of measurements, are close to 100% recognition rate (Web-1), (Pittman, 2007), offline methods are far from acceptable performance. If we consider the processing of archive documents, the problem is more difficult, since there are several phenomena making the appearance of words noisy and/or distorted. The transparency of papers, overlapping words, dust, blobs make most recognition methods to fail.
Manual processing of such archive documents is not feasible since the number of pages only in Hungary is approximately 3,500.0 million of which 200 million is recommended for digitization (Orosz et al., 2008).

In (Czuni et al., 2013) a SIFT (Scale Invariant Feature Transform, (Lowe, 1999)) based word spotting approach was proposed to recognize handwritten words of archive documents. The motivation was based on observations that lots of archive documents have a very limited vocabulary (see Czuni et al., (2013)). This observation might be valid for several types of documents such as juridical, governmental, or medical texts. Our testbed consists of 22 pages from a book of census of a medieval city. The book from the 1770's has a few thousand expressions with a few hundred classes of expressions, family and Christian names almost written in cursive style. The layout of the book is shown by Fig. 1. (left). The test database contained 1638 manually segmented and annotated words of 177 different names. The average occurrence of a word is 9.25 (ranging from 1 to 111). As described later, the word spotting method applied could reach about

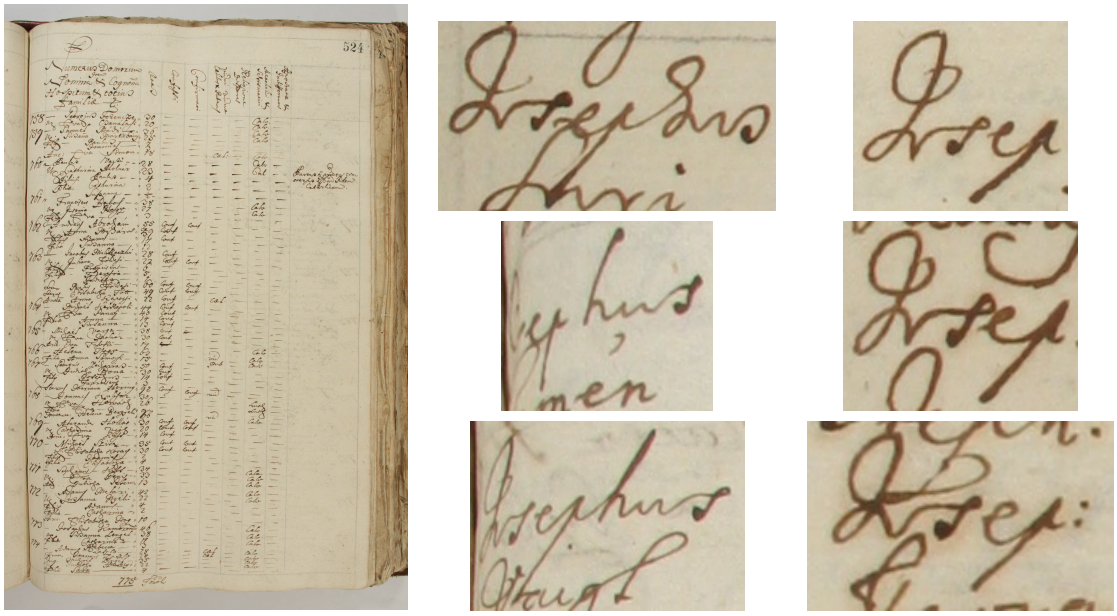

Fig. 1. A typical page of the archive document under study and some examples to illustrate the difficulty of the task (different forms of Josephus and Josep suffering distortions and overlapping).

75-84% hit rate (depending on the test procedure). This can be considered as a fair result considering the level of difficulty. Fig. 1. (right) illustrates some of the typical problems of the archive documents under study. However, beside the recognition rate computational complexity is also a problem to be considered. One typical task is to find a specific word in a document (which is basically a 1:n search in word spotting of document size n). If the task is the recognition of all words of a document it means an m:n search type with m $\times$ n comparisons (of words of different shape). This large number of comparisons makes the spotting itself practically unfeasible without additional classification or fast searching techniques. Unfortunately typical classification or Bag of Words techniques easily fail due to the the large variability of shape and the rare occurrences of many of the word classes. Our paper focuses on this problem and proposes the use of a pivot image set as reference objects used during the search.

2. Related Works

It is very common that handwriting recognition needs preprocessing to get rid of such problems as overlapping, the slant of letters, blotches, or the imprint of words from the backside of the paper. For example (Lavrenko et al., 2004) first manually separates words to remove parts from overlapping, then the slant of words and skew of rows are estimated and compensated. As a third step the background colour of the word is normalized and the baseline of the word is positioned. After preprocessing, pixel domain features (e.g. the number of descenders and ascenders, width, height, etc.) and Discrete Fourier Transform coefficients are used in a Hidden Markov Models (HMM) to achieve about 65% accuracy on some historical documents. (Rath, Manmatha, 2007) also uses preprocessing to remove skew of rows and slant of words. Comparison of word images is based on the upper and lower profiles and the number of foreground–background transitions. The best minimal error rate was around 31.5% on some archive documents. Dynamic Time Warping (DTW) was also involved to compensate horizontal distortions. Our basic comparison technique is similar, in some sense, to the method of (Rothfeder et al., 2003), where it is proposed to recover correspondences of feature points. These were detected by the Harris corner detector, and then correspondences are used to construct a

similarity measure of two images of the two words. The Sum of Squared Differences (SSD) error measure was applied to compare gray-level intensity windows, which are centered on detected corner locations. The Euclidean distance of these locations was used to rank word images in the order of their closeness to a query image. (Rodrigez, Perronnin, 2008) uses local gradient features for word comparisons. Gradient histograms are calculated in a moving window and the generated feature vectors are compared either by DTW or by HMMs. The equal error rate was found between 12% and 30%, depending on the exact method, but the technique was tested only with the 10 most frequent word classes of the database. Please note, that this method can be considered such as a simplified SIFT, however it does not have rotational and scale invariance. In (Zhang et al., 2009) another method similar to SIFT was applied for the recognition of Chinese characters. While authors state that they did not found the original SIFT features stable enough in case of different writing styles, this is not supported by test data. Instead, they apply elastic meshing and dynamic gradient histogram calculation without scale or rotation invariance. (Fischer et al., 2010) first transforms the images into binary form then uses normalization of skew and scale. The skeleton of these normalized binary objects is processed to code the keypoints (endpoints, intersections and corner points of circular structures) in graphs. Graphs are compared by graph edit distance and HMMs are also utilized. Results seem to be good for the Parzival data set (a manuscript from the 13th century). We should note that these images do not contain cursive (joined-up) letters but the so called *fraktur* (broken) script style. In (Kluzner et al., 2009) first image distortion compensation is done by a modified optical flow method. Once the pixel displacement vector is computed, compensation of image is performed similarly to motion compensation in video coding. They binarize both images (query and candidate) and the difference between the two binary images is computed using a non-linear difference measure. Test results are above 80% on a database of 18th century old fraktur German Gothic fonts. In (Uchida, Liwicki, 2010) the Speeded-up Robust Features (SURF) (Bay et al., 2008) descriptor is applied to recognize handwritten digits. Several feature points per digit are described by 128-dimensional vectors and classified into one of the ten possible classes. The position of SURF points is disregarded. Then at the character level each feature point votes for one of the classes so the classification is solved by counting the votes. Higher than 90% recognition rates are achieved on the 10 numeric characters. In (Rothacker, 2013) a combination of HMMs and Bag of Words was utilized to find occurrences of words in an archive document page (basically a 1:n search an opposite task of ours). The method was tested on the George Washington dataset with good results. While the complexity of the method is not discussed the precision was high (about 61%) and the method did not require the segmentation of the document. In the following sections we will discuss a word spotting approach where the similarity of two words is based on the matching of their SIFT points. The main advantage of the technique is the ability to find corresponding words in case of large distortions and noise while preprocessing of the images is not necessary. Analysis of the technique is in (Czuni et al., 2013), now after the brief description of the core algorithm we will propose a pivot based search for the effective speed up of the aproach.

3. The Proposed Method

3.1. The Similarity of Words

Our proposed method is based on the similarity comparisons of two images represented by their SIFT points. To recognize a word we must find the most similar image from a previously learnt set of labelled words. The similarity value for the query (Q) and candidate (C) words:

$$S(Q,C) = \sum_{j=1}^{N} (\sqrt{255^2 \cdot 128} - D(q_j, c_j)_{(q_j,c_j)\in M_{Q,C}}), \quad (1)$$

where $M_{Q,C} = \{(q_i, c_{i,min}), i = 1,2,...,N\}$ is the set of matching points of size N. $D(q_i, c_j)$ is the distance of these points:

$$D(q_i, c_j) = \sqrt{\sum_{k=1}^{128} (q_i(k) - c_j(k))^2}. \quad (2)$$

where $q_i(k)$ and $c_j(k)$ are the SIFT components of size 128. For each SIFT point q_i of the query image Q we find the most similar candidate point:

$$c_{i,min} = \min_{c_j} D(q_i, c_j) \quad (3)$$

in a limited spatial distance (within a circle) and at similar orientation. For the detailed description of the similarity estimation please see (Czuni et al., 2013).

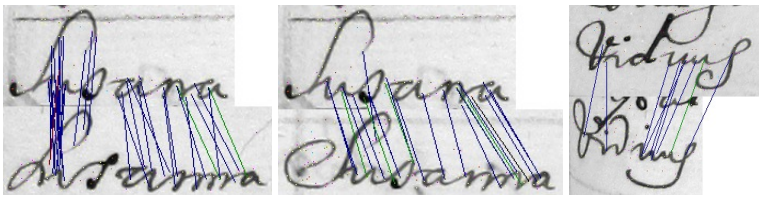

Fig. 2. Examples for the matching of SIFT points of Susanna and Vidius.

We note at this point that there is no need for preprocessing (e.g. binarization, slant correction, noise removal) and precise segmentation due to the scale and rotation invariance of SIFT and the nature of the matching mechanism. Also DTW can be set aside since the applied searching area (spatial searching circle) allows horizontal (and even vertical) distortion of the shapes. In the tests presented we used only a rough (manual) segmentation of words, the automatic segmentation is out of the scope of our paper. Although a full search ends up in quite good retrieval performance (75-85% hit rate on our documents of about 1500 words), the main disadvantage of this approach is the high computational cost. First, the calculation of SIFT is intensive. Second, we have to compare the query to all of the possible candidates. Third, since $S(Q,C)$ is not symmetric, thus it is to compute both $S(Q,C)$ and $S(C,Q)$ and make the average of the two: $S_{Sym} = (S(Q,C) + S(C,Q))/2$.

3.2. Using Pivots as Reference Words

The idea of using reference patterns (e.g. basis images, eigen images) for image comparisons is a well-known technique in image processing. Karhunen-Loewe Transformation (or the equivalent Principal Component Analysis), and Fourier Transformation are good examples when image data are described and compared with the help of only a subset of all transformation coefficients. There are two advantages of these methods: they help to eliminate noise and they can also reduce the dimension of data vectors. Unfortunately, the image of handwritten text is like a line drawing affected with nonlinear geometrical transformations, thus the generation and usage of eigen-images and corresponding eigen values would not result in effective representation. Instead, we propose to select some of the training words as pivots which can serve as reference images. This way we can transform the problem from the non-metric space of similarity values S_{Sym} to similarity descriptors where correlation based metric will show efficiency. The main points of our method are the following:

1. Select T training words. All possible classes should have at least one appearance in this database of training words.

2. Calculate the similarity (S_{Sym}) of all these words generating a TxT symmetrical similarity matrix. (The lines of this matrix are centralized by the subtraction of the average of each line.)

3. Select pivot words from the training examples, forming the pivot set S_p. The method of selection is discussed later. To achieve speed-up the number of pivots should be significantly smaller than T.

4. To find the word most similar to a query image compare it to the pivot words and use the similarity values as a vector to compare to all other words.

Thus instead of computing the similarity of a query word to all possible candidate words, we only compute $S_{Sym}(Q,C_i)$ to the pivot words ($i \in S_p$) to generate the *Pivot Similarity Descriptor vector - PSD*(Q) of dimension T. The obtained *PSD* vector can be used to compare a query to all possible images previously shown. For such comparisons we can use the correlation of *PSD* vectors of queries *PSD*(Q) and candidates *PSD*(C_i).

3.3. Selection of Pivots

The reader may ask why not use the similarity values of words, as descriptors, for classification. We tested Support Vector Machines (with Radial Basis Function kernel ($RBF_\gamma = 0.003$) and regularization $C = 7$) to classify words but the best classification rate was about 65% while other classifiers were even worse.
The main advantage of our approach is that we don't apply classification at all so can avoid the loss of rare or outlier object instances. This is very important in example based word spotting. We tested two approaches for pivot selection:

1. Selecting the top n pivot words based on words similarity values importance. Since words similarity values are continuous against a categorical target, importance is ranked using the F statistics.

2. Selecting the top n pivot words based on SVM classifier predictor importance.

The first selection ranks each similarity value based on the strength of its relationship to the specified target, independent of other inputs, the SVM predictor importance indicates the relative importance of each input. It is quite obvious to see that the second solution easily outperforms the other as shown in the next section.

4. Tests and Analysis

For testing the pivot based search we used a testbed of 1638 test images. The query images were excluded from the candidates in all cases. Fig. 3. (left) shows the hit rate as the function of number of pivots. It is important to mention that each query had at least one representative in the candidate set. That is there is no sense of talking about *negative* queries and counting *true-* or *false negative* results but *hit rate* (also called *recall*) are satisfactory in evaluations.

We also investigated the probability of the good results falling in the set of the best N retrieved result. Fig. 3. (right) shows the results for $N = 10$ and Fig. 4. shows the hit rate in the first N retrieved images.

4.1. Two or Three Phases Search

As described above now the search basically consists of two phases:

1. Compare the query to the pivots to get its PSD vector.

2. Compare the PSD to the PSDs of candidates with centralized correlation.

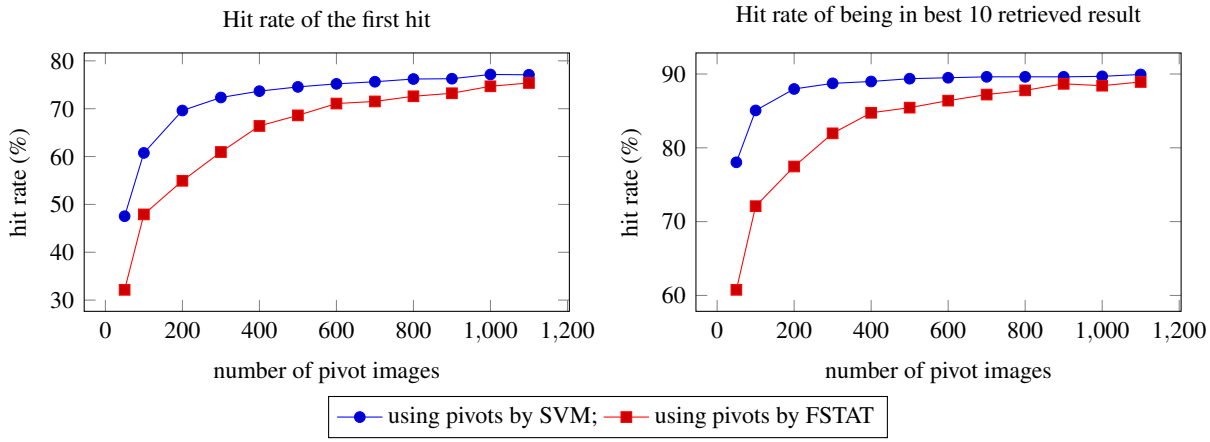

Fig. 3. Hit rate as the function of number of pivots and hit rate of being in the best 10 retrieved result as the function of number of pivots.

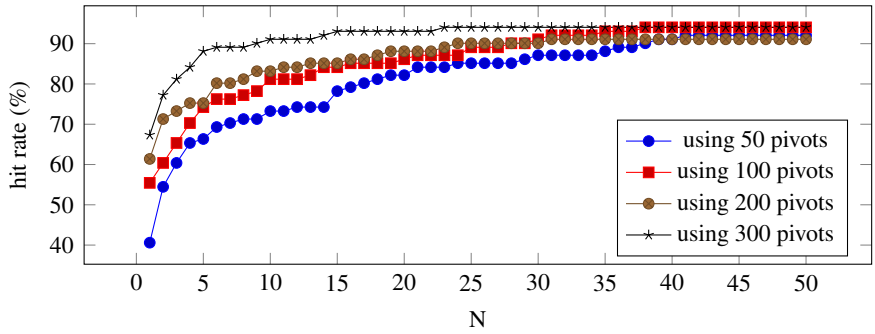

Fig. 4. Hit rate of being in the top N results in case of different number of pivot points.

The computational cost of the first step is determined by the calculation of $S(Q,C)$ (see e.q. 1) while the second step depends on the size of the set of candidates. Considering Fig. 4 we can introduce a third phase, where we apply a linear search on the best N candidates obtained in the second step. That is the number of $S(Q,C)$ calculations equals the number of pivots plus N. Fig. 5 shows that a slightly better performance can be achieved this way at around 78% hit rate.

5. Summary and Future Works

In this paper we dealt with word spotting for the recognition of handwritten archive text. We described two basic problems to be solved: to find a good similarity measure of heavily distorted word images and to build a fast search algorithm for the matching of thousands of example images. It has been shown that the proposed SIFT based similarity measure can be used efficiently for the first problem with a conjunction of a pivot based search in the similarity space.

In our testbed we could achieve about 78% hit rate using 300 pivots. The speedup of pivot based search is about 2.4 times in case of an example vocabulary of size 1600 while about 3.5 times at a vocabulary size of 4000. The average running time for a query on Intel(R) Core(TM) i7-3610QM @2.3 GHZ CPU is illustrated on Fig. 6. (right). A possible application of the method is by archivists or historians where semantic word

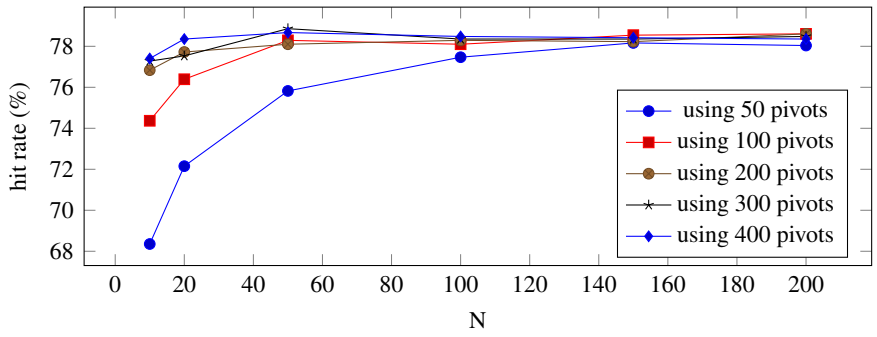

Fig. 5. Hit rate of three-phase search in case of different number of pivot points.

classes are used (for example Josep and Josephus are within the same class). Evaluation of the method in such a context results in a slightly better performance as illustrated in Fig. 6. (left). For further speedup we are planning to make clustering of words in PSD space. Also in future we plan to test our approach on other datasets.

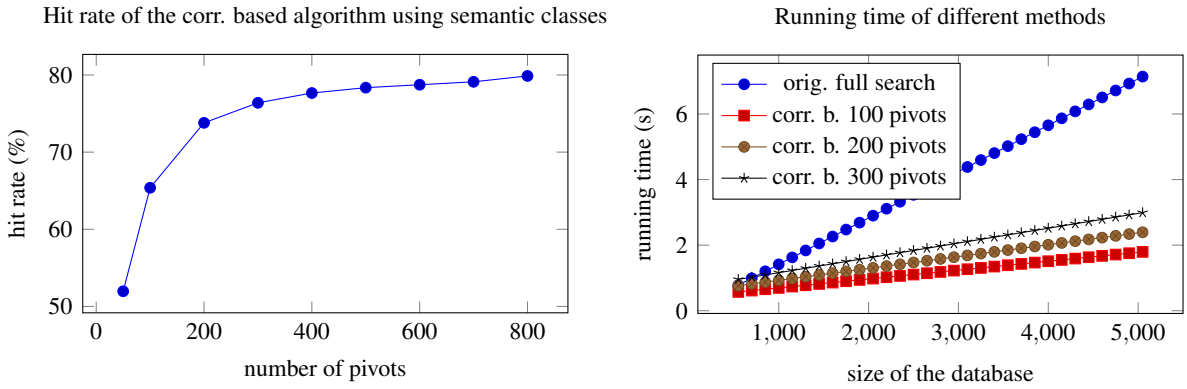

Fig. 6. Hit rate of semantic classes as the function of number of pivots and the running time of one query using different methods

Acknowledgments

The work and publication of results have been supported by TÁMOP–4.2.2.C-11/1/KONV-2012-0004, and by the Bolyai scholarship of the Hungarian Academy of Sciences.

References

Bay, Herbert; Ess, Andreas; Tuytelaars, Tinne and Gool, Luc Van *Speeded-Up Robust Features (SURF)*, Comput. Vis. Image Underst., 110, 3, 346–359 (2008)

Czúni, László; Kiss, Péter József; Gál, Mónika; Lipovits, Ágnes Local Feature Based Word Spotting in Handwritten Archive Documents, 2013 11th International Workshop on Content Based Multimedia Indexing 17 - 19. June 2013

Fischer, A.; Riesen, K.; Bunke, H., *Graph Similarity Features for HMM-Based Handwriting Recognition*

in Historical Documents, Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on, 253–258 (2010)

Kluzner, Vladimir; Tzadok, Asaf; Shimony, Yuval; Walach, Eugene; Antonacopoulos, Apostolos *Word-Based Adaptive OCR for Historical Books*, Int Conf. on Document Analysis and Recognition, 501–505 (2009)

Lavrenko, V.; Rath, T. M.; Manmatha, R. *Holistic Word Recognition for Handwritten Historical Documents*, International Workshop on Document Image Analysis for Libraries, 278–287 (2004)

Lowe, D. G. *Object Recognition from Local Scale-invariant Features*, Proceedings of the International Conference on Computer Vision 2, 1150–1157 (1999)

Orosz, Katalin; Rácz, Gyrgy; Reisz, T. Csaba; Vajk, Ádám; Véber, János; Közpkori oklevelek tömeges digitalizálása, Magyar Országos Levéltár, (2008)

Pittman, J.A.: Handwriting recognition: tablet PC text input. IEEE Comput. 40(9), 49-54 (2007)

Rath, T. M.; Manmatha, R. *Word Spotting for Historical Documents*, Int. Journal on Document Analysis and Recognition, 139–152 (2007)

Rodríguez, José A., and Florent Perronnin, *Local Gradient Histogram Features for Word Spotting in Unconstrained Handwritten Documents*, Int. Conf. on Frontiers in Handwriting Recognition, 7–12 (2008)

Rothacker, L.; Rusinol, M.; Fink, G.A., "Bag-of-Features HMMs for Segmentation-Free Word Spotting in Handwritten Documents," Document Analysis and Recognition (ICDAR), 2013 12th International Conference on , vol., no., pp.1305,1309, 25-28 Aug. 2013

Rothfeder, Jamie L.; Feng, S.; Rath, Toni M., *Using Corner Feature Correspondences to Rank Word Images by Similarity*, Computer Vision and Pattern Recognition Workshop, CVPRW '03. Conference on , vol.3, 30–30 (2003)

Uchida, S.; Liwicki, M., *Part-Based Recognition of Handwritten Characters*, Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on, 545–550 (2010)

Zhang, Zhiyi; Jin, Lianwen; Kai Ding; Xue Gao, *Character-SIFT: A Novel Feature for Offline Handwritten Chinese Character Recognition*, Document Analysis and Recognition, ICDAR '09. 10th International Conference on, 763–767 (2009)

Website References

http://www.cse.ust.hk/svc2004 visited: 14 March 2014