Pattern Recognition 48 (2015) 3346–3356

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Handwritten word image matching based on Heat Kernel Signature

Xi Zhang *, Chew Lim Tan

School of Computing, National University of Singapore, 13 Computing Drive, 117417 Singapore

ARTICLE INFO

Article history:
Received 1 February 2014
Received in revised form
16 October 2014
Accepted 21 October 2014
Available online 28 October 2014

Keywords:
Handwritten documents
Word image matching
Heat Kernel Signature
Triangular mesh
Matching score

ABSTRACT

Keyword spotting is an alternative method for retrieving user specified text or query word images, without Optical Character Recognition (OCR), by calculating the similarity between features of word images rather than ASCII content. However, because of unconstrained writing styles with large variations, the retrieving results are always not very satisfactory. In this paper, we propose a novel method, which is based on Heat Kernel Signature (HKS) and triangular mesh structure to achieve handwritten word image matching. HKS can tolerate large variations in handwritten word images and capture local features. On the other hand, the triangular mesh structure is used to present global characteristics. Moreover, our method does not need pre-processing steps.

1. Introduction

A large amount of valuable handwritten documents, such as business contracts, historical manuscripts, application forms, diagnosis reports, and envelops, have been scanned into digital databases or libraries and public access services are provided for users to retrieve useful information they need. In order to achieve efficient and reliable retrieving services, OCR (Optical Character Recognition) is first used to convert image-based documents into ASCII format. However, degradation, noise and various unconstrained writing styles always prevent OCR providing satisfactory recognition results. Keyword spotting [1] becomes an alternative way of OCR to retrieve information for users, especially it can be used for spotting query text or word images. Furthermore, only features extracted from the query images are needed, and the ASCII content can be unknown.

A method which is commonly used to achieve handwritten word image matching is extracting geometrical features in each column of the word images from left to right [2] and applying Dynamic Time Warping (DTW) [3] to calculate the distance between two sequences of feature vectors. However, pre-processing steps are needed and very crucial, including binarization, skew or slant correction, and normalization, therefore the accuracy highly depends on the reliable pre-processing results. Moreover, column features only take the current column into account and ignore the context information. Consequently, DTW based on column feature sequences may not deal with word images with large variations, which are always the cases in handwritten documents. In order to consider context information of the consecutive strokes in handwritten words, [4] extracted features from a sliding window instead of only one column.

Some other widely used methods are based on Scale Invariant Feature Transform (SIFT) [5], which has been successfully used in computer vision and object recognition, and also shows its robustness and reliability to be invariant to multi-scaling and -rotation of images. Some variations of SIFT feature are gradually used for document analysis. In [6], a new feature sequence is proposed based on the local gradient histogram based on the idea of SIFT, which is extracted from each cell of a sliding window. However, this method also depends on pre-processing steps.

In handwritten documents, infinite writing styles may occur, so that different writers, or even the same writer, may write the same word in large variant styles, just like the same word image is deformed by non-rigid deformations in every part of the strokes. Moreno-Noguer [7] shows that SIFT can deal with affine-invariant situations quite well, but cannot handle non-rigid deformations. On the other hand, Heat Kernel Signature (HKS) [8] is proved to be invariant to non-rigid deformations. Motivated by this observation, we propose a new method for handwritten word image matching based on HKS. We also propose a new similarity measurement approach to calculate the distance between two sets of descriptors, based on triangular mesh structure, which can capture global spatial relations of keypoints. Moreover, our method do not need pre-processing steps, such as binarization, normalization, and skew or slant correction.

In this paper, we will first introduce the localization of keypoints and how the HKS is extracted from each keypoint in a word image, which will construct the descriptor for the word image. Then, we will present our new method for computing the distance

* Corresponding author.
E-mail addresses: xizhang@comp.nus.edu.sg (X. Zhang), tancl@comp.nus.edu.sg (C.L. Tan).

http://dx.doi.org/10.1016/j.patcog.2014.10.028
0031-3203/

between two descriptors of word images, based on the triangular mesh structure and score matrix. Finally, we will test our method on large variant handwritten word images.

2. Descriptor based on Heat Kernel Signature

2.1. Keypoints detection and selection

When HKS is used for shape segmentation or recognition, all the points on the surfaces are used to measure the distances between two shapes, however, for handwritten documents, generating descriptors for all the points in the word images [9] and finding the optimal matching between two sets of descriptors are very time consuming and unnecessary [10]. In order to localize informative keypoints in the word image, we apply the keypoint detector for SIFT proposed by Lowe [5].

In handwritten documents, we focus on the keypoints located on the strokes, namely, we remove the keypoints in the background or near the contour of strokes. Therefore, word images are first smoothed by a Gaussian kernel, and keypoints of each word image are located, such as the red points shown in Fig. 1(a). Then, the keypoints with the intensity value smaller than a threshold are removed, which are always located in the background. Fig. 1(b) shows the final selected keypoints, the descriptors of which will be generated, as described in Section 2.2.

2.2. Heat Kernel Signature

Heat Kernel Signature (HKS) was first proposed for 3D shape recognition or classification. Based on the 3D coordinates of all the points on the shape surface and their triangular mesh structure, heat kernel can capture the characteristics of the shape with the Laplace–Beltrami operator. The heat kernel is an isometric invariant, due to the invariance property of the Laplace–Beltrami operator. Therefore, the heat kernel can even match the same human or animal with different poses. Moreover, due to its multi-scale property, we can capture the features in its near neighborhoods or on the global shape, so that if two points have similar features in their small neighborhoods, but they may have very different features on the whole shape. Therefore, we can match two points by the features from their small neighborhoods to large domain. In the rest of this section, we will introduce how to calculate HKS descriptors.

After the keypoints are detected in the word images, HKS descriptor will be extracted from a local patch centered at each keypoint. Heat Kernel Signature can capture local geometry of 3D shapes with short time scales and gradually represent global characteristics as time becoming larger [7]. The heat kernel has the properties of invariant to isomeric, and stable to non-rigid deformations. Moreover, it can capture both local and global characteristics.

In order to apply HKS to word images, we should first embed a patch in a 2D word image into a 3D surface. We assume that P is a patch, with the size of $N \times N$, extracted from a word image I and centered at a keypoint, as shown in Fig. 2(a). The 3D Riemannian manifold M is the 2D surface embedded into 3D space from P, satisfying the condition that if (x,y) is a point in P, then there is a point (x,y,z) on M, where z is the intensity value of (x,y) in P [7], as shown in Fig. 2(b). The heat diffusion geometry of patch P is obtained by using the Laplace–Beltrami operator over the manifold M [11]:

$$\left(\Delta_M + \frac{\partial}{\partial t}\right) u(\mathbf{x}, t) = 0, \tag{1}$$

where $\mathbf{x}$ is a point on M, Δ_M is the Laplace–Beltrami operator, and the solution $k(\mathbf{x}, \mathbf{x}', t)$ is named as the heat kernel, presenting how the heat between two points $\mathbf{x}$ and $\mathbf{x}'$ on the same surface diffuses from one to the other at time t, if we assume that the unit heat source is from the position of $\mathbf{x}$ at time $t=0$. When M is a compact manifold, $k(\mathbf{x}, \mathbf{x}', t)$ can be expressed compactly by the eigenvalues $\{\lambda_i\}$ and eigenvectors $\{\phi_i\}$ of Δ_M [11]:

$$k(\mathbf{x}, \mathbf{x}', t) = \sum_{i=0}^{\infty} e^{-\lambda_i t} \phi_i(\mathbf{x}) \phi_i(\mathbf{x}'). \tag{2}$$

Based on Eq. (2), HKS is proposed in [12] to present local and global characteristics around a point $\mathbf{p}$ on M as follows:

$$\mathrm{HKS}(\mathbf{p}, t) = k(\mathbf{p}, \mathbf{p}, t) = \sum_{i=0}^{\infty} e^{-\lambda_i t} \phi_i^2(\mathbf{p}). \tag{3}$$

In order to tolerate 2D noise around keypoints [7], the descriptor of $\mathbf{p}$ is constructed from all the points in P, weighted by a Gaussian kernel considering their distances to the center. This descriptor is called Deformation Invariant (DI) descriptor [7]:

$$\mathrm{DI}(\mathbf{p}, t) = [\mathrm{HKS}(\mathbf{x}, t) .\!* G(\mathbf{x}; \mathbf{p}, \sigma)]_{\forall \mathbf{x} \in \mathbf{P}}, \tag{4}$$

where G is a 2D Gaussian filter centered at p with standard deviation σ. Fig. 3 shows the DI descriptors of the patch in Fig. 2 (a) at different t values, where the number of eigenvalues and eigenvectors we use in Eq. (3) is 100.

2.3. Discrete version of the Laplace–Beltrami operator

Because we have finite number of points in a patch, we apply the discrete version of the Laplace–Beltrami operator based on cotangent scheme [13], in order to obtain $\{\lambda_i\}$ and $\{\phi_i\}$.

We assume that $P = \{p_1, p_2, \ldots, p_v\}$ are all the pixels in a patch, and the intra-pixels as shown in Fig. 4. We construct the triangular mesh based on P. The discrete version of the Laplacian matrix L is a $v \times v$ matrix, and is computed as follows [8]:

$$m_{ij} = \begin{cases} \dfrac{\cot\alpha_{ij} + \cot\beta_{ij}}{2} & \text{if } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$L_{ij} = \begin{cases} \sum_k \dfrac{m_{ik}}{s_i} & \text{if } i = j \\ -\dfrac{m_{ij}}{s_i} & \text{if } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where α_{ij} and β_{ij} are the angles depicted in Fig. 4(c), s_i is the area of all the triangles having the same vertex p_i. In order to compute the eigenvalues and eigenvectors of the non-symmetric matrix L, let S be a diagonal matrix with $S_{ii} = s_i$, and M with $M_{ij} = m_{ij}$, so that $L = S^{-1}M$. The generalized eigenvalue problem of L can be written as

$$M\vec{v} = \lambda S \vec{v} \tag{7}$$

where λ and $\vec{v}$ correspond to $\{\lambda_i\}$ and $\{\phi_i\}$ in Eq. (3). The computational time cost for HKS is always very high, especially

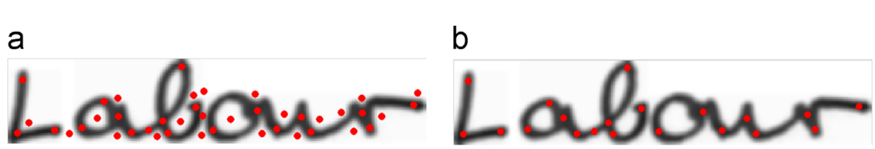

Fig. 1. Keypoints selection and the corresponding triangular mesh structure. (a) Keypoints detected by SIFT detector. (b) Removing uninteresting points. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)