Pattern Recognition Letters 43 (2014) 119–126

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Structural similarity for document image classification and retrieval

Jayant Kumar *, Peng Ye, David Doermann

Language and Media Processing Laboratory, Institute of Advanced Computer Studies, University of Maryland, College Park, United States

ARTICLE INFO

Article history:
Available online 12 November 2013

Communicated by Katsushi Ikeuchi

Keywords:
Structural similarity
Retrieval
Classification
Random forest

ABSTRACT

This paper presents a novel approach to defining document image structural similarity for the applications of classification and retrieval. We first build a codebook of SURF descriptors extracted from a set of representative training images. We then encode each document and model the spatial relationships between them by recursively partitioning the image and computing histograms of codewords in each partition. A random forest classifier is trained with the resulting features, and used for classification and retrieval. We demonstrate the effectiveness of our approach on table and tax form retrieval, and show that the proposed method outperforms previous approaches even when the training data is limited.

1. Introduction

Finding structurally similar images in large heterogenous document image collections has been of interest for many years (Shin and Doermann, 2006; Marinai et al., 2011). While there are numerous applications in office automation, litigation support and general document image search which depend on efficient and effective methods for computing similarity, previous approaches have focused on content-specific features or layout-specific structures (Collins-Thompson and Nickolov, 2002; Shin and Doermann, 2006; Zhu et al., 2009). Approaches based on content are highly dependent on, and sensitive to, the quality of optical character recognition (OCR), graphics recognition or component labeling. Since the OCR for unconstrained handwritten documents is still a difficult problem, content based approaches are typically limited to more structured machine printed documents (Marinai et al., 2011). Furthermore, layout based approaches tend to be tailored to fixed layouts, and model known classes of documents such as articles or forms. There is however an emerging need for effective methods for unconstrained document images for which OCR cannot be performed.

Recent work has focused on developing general methods capable of handling less constrained handwritten documents and datasets with highly variable layout (Kumar et al., 2011, 2012; Jain and Doermann, 2012). Moreover, approaches which move beyond fixed partitions and compute similarity at different levels can adapt by allowing the user to specify the degree of similarity. For example, a range from 0.0 (no match) to 0.5 (conceptual match) to 1.0 (exact match).

Structural similarity therefore becomes important when users want to supplement search for images using visual content like logos, signatures, tables, etc., with search for *layout characteristics*. In such cases, users may or may not fully understand the layout or structural characteristics they are interested in, so they can either provide a sketch (or explanation) or provide some *representative* documents as examples. So they do not have to make these characteristics explicit, it becomes important to capture similarity at various levels, from the low-level content to high-level structure. Approaches developed for content-based matching and retrieval alone cannot be directly applied as they lack a high-level representation.

One effective way to define layout similarity for matching is based on structural features (Collins-Thompson and Nickolov, 2002; Shin and Doermann, 2006; Joutel et al., 2007). However, hand-crafting structure-based features (e.g., spatial relationships among the components) in unconstrained and noisy documents is difficult due to variation in content, translation, rotation and scale of components. Furthermore, as previously mentioned, a majority of the work published on defining and applying structural similarity is specific to a particular document type, such as business letters (Dengel and Dubiel, 1995; Marinai et al., 2006). The problem is made even more difficult when the number of relevant images for training is limited (Zheng et al., 2004).

In this work, we present a method for the classification of structurally similar document images which can be applied to a broad class of documents. By *structural* similarity we mean primarily the layout and spatial organization of document content, including text, signatures, lines, logos, table-elements, etc. in documents. Structurally consistent match between two document images is a match that preserves the constraints of one-to-one mapping and parallel connectivity (Forbus et al., 1995). One to one mapping requires that for each element in one image there exist a similar

* Corresponding author. Tel.: +1 2406015180.
E-mail addresses: jayant@umiacs.umd.edu (J. Kumar), pengye@umiacs.umd.edu (P. Ye), doermann@umiacs.umd.edu (D. Doermann).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.10.030

corresponding element in other image. For example, if there is title or heading in the document, centered at the top, then having a similar heading at the same place in another document will be considered an exact match; having a title with different text will be consider approximate match, and if there is completely different element such as a figure, signature or logo, then that is not a match.

It is useful to map structural similarity to a scale from 0 to 1 where higher values indicates more precise match between document objects. Of course the similarity above which two documents are considered in same class depends on a specific application. A tax-form and a bank-form for example are structurally similar if we are interested in form retrieval, but are dissimilar if we are interested in retrieving a specific instance of a form.

Our approach is based on statistics of robust local features in different partitions of an image. The structure and layout of document objects such as text-lines, margins in text-blocks, lines in tables and border-designs typically run across both horizontal and vertical directions (Fig. 1). To capture spatial relationships and correlations, we recursively divide the image horizontally and vertically, and compute histograms of *learned* codewords in these regions. We show that this strategy of modeling spatial relationships results in increased accuracy using the random forest (RF) classifier, even when only a few labeled samples are used for training.

In Kumar et al. (2012), we explored an unsupervised feature learning method, using raw-image patches, to construct a codebook representation of basic structural elements in document images. Since raw-image patches are not scale-invariant and are less robust to noise present in the monochromatic images, it required a large codebook to achieve good performance. In this work, we extend that approach. First, we use SURF features as a basic unit of local content. SURF descriptors are more robust to noise and are scale-invariant. Second, we show that the approach is effective for *in-class* table and tax-form discrimination requiring very few labeled samples for training, and present classification results on 53 classes of hand-drawn table images and 20 classes of tax-form images. We compare our approach with the spatial-pyramid method (Lazebnik et al., 2006) and show that the proposed method gives superior performance on many document retrieval and classification tasks.

The remainder of this paper is organized as follows. In Section 2 we present related work on the retrieval of structurally similar document images. We discuss the details of our approach in Section 3. We present experimental results in Section 4 and conclude the paper in Section 5.

2. Related work

There are a number of paradigms in which document image retrieval can be performed. For text-content based retrieval, scanned document images are typically converted to electronic (Unicode) text through optical character recognition (OCR) (Decurtins and Chen, 1995). More recent retrieval approaches have focused on image-based representations allowing a focus on visual representation. When considering layout, the representation of documents using image-based features is often more intuitive and useful because it preserves the physical structure and access to non-text components such as embedded graphics (Marinai et al., 2011).

A large number of retrieval techniques have been developed using a query by example paradigm (Zhu and Doermann, 2009; Marinai et al., 2011; Jain and Doermann, 2012; Chen et al., 2012a), where features are extracted and indexed from document images off-line. A query image (e.g., words, logos, signatures) is provided, and features are extracted and matched against the indexed database of features. Documents which result in a number of matches above a certain threshold are considered relevant and can be geometrically verified (Zhu and Doermann, 2009; Jain and Doermann, 2012). All these works emphasize the importance of using robust and scale-invariant descriptors for matching.

An alternative approach defines *similarity* based on the *model* trained using features (possibly class specific) extracted from a user-provided set of *example* documents. Shin and Doermann (2006) defined *visual similarity* of layout structures and applied supervised classification for each specific type. They used image features such as the percentage of text and non-text (graphics, images, tables, and rulings) in content regions, column structures, relative point sizes of fonts, density of content area, and statistics of features of connected components. They used a decision tree classifier and self-organizing maps for classification. The main drawback of their approach is that the features were designed for specific document classes (e.g., forms, letters, articles). Additionally, due to a large number of different feature types the approach is computationally slow for large scale document exploration.

Collins-Thompson and Nickolov (2002) proposed a model for estimating the inter-page similarity in ordered collections of document images. They used features based on a combination of text and layout features, document structure, and topic concepts to discriminate between related and unrelated pages. Since the text from OCR may contain errors, especially for handwritten documents, the approach is limited to well-structured printed documents. Joutel et al. (2007) presented an approach for the retrieval of handwritten historical documents at page level based on the *curvelet transform* to compose a unique signature for each page. The approach is effective when local shapes are important for classification but the approach is likely to miss any higher level of structural saliency. In many cases, the desired similarity is embedded in global structure and relationships among different objects in document images. In our approach, similarity is computed at two levels: first, a local match is performed using SURF based codewords and second, statistics of different codewords in different partitions are considered for higher level structure match.

Approaches based on bag-of-words (BOW) models have shown promising results on many computer vision tasks such as image classification (Wallraven et al., 2003), scene understanding (Quelhas et al., 2005), and document image categorization (Barbu et al., 2006; Kumar et al., 2011). However, initial formulations for computing similarity typically disregard the spatial relationships between codewords, and only consider the occurrences of each codeword in an image. This results in a limited descriptive ability and performance degrades in presence of noise, background

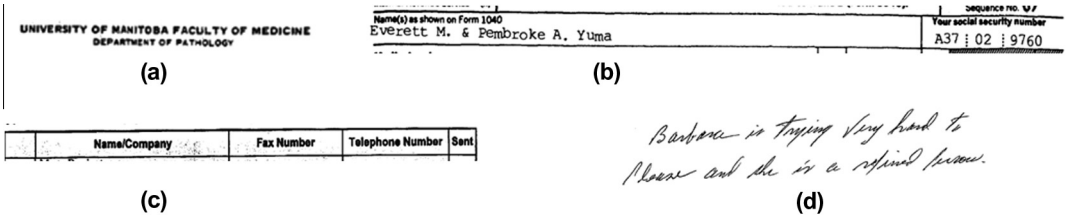

Fig. 1. Document objects from Tobacco database showing horizontal bias.