An Efficient Method for Extracting Family Records from Mixed-type Forms

Xujun Peng
Raytheon BBN Technologies
10 Moulton St, Cambridge, MA 01801
xpeng@bbn.com

Huaigu Cao
Raytheon BBN Technologies
10 Moulton St, Cambridge, MA 01801
hcao@bbn.com

Krishna Subramanian
Raytheon BBN Technologies
10 Moulton St, Cambridge, MA 01801
ksubrama@bbn.com

Rohit Prasad
Raytheon BBN Technologies
10 Moulton St, Cambridge, MA 01801
rprasad@bbn.com

Prem Natarajan
Raytheon BBN Technologies
10 Moulton St, Cambridge, MA 01801
pnataraj@bbn.com

ABSTRACT

In this paper, we propose a framework for extracting information of interest from forms that contain both machine-printed and handwritten text. The main challenge of data extraction from such documents is in accurately localizing the text regions of interest, and then recognizing the content in these regions. In this paper, we present a novel template matching method that uses Affine Scale-invariant Feature Transformation (ASIFT) to extract key points in a test image and pre-specified templates. Our experimental results show that the proposed method can effectively and accurately locate and extract the data entries for mixed-type forms.

General Terms

Application

Keywords

Document, Extraction, ASIFT

1. INTRODUCTION

While plenty of information in documents is now available in digital format which can be easily accessed, a large amount of historical paper documents are mostly unexplored. In particular, a capability for automatically extracting information from mixed-type forms that contain both printed and handwritten text such as census and tax records can be extremely useful for a wide variety of applications. Unlike free-style handwritten text, forms have a well-defined structure that provides useful cues for document classification, registration, and data extraction. Form structure has been exploited for several tasks in prior work in this area. For instance, to classify different type of tax forms, Xu et.al designed a hierarchical classifier to classify the forms according to their layouts by using line and key point features [6]. In the work of [5], Sako et. al used a key word template matching method to identify tax forms and extract region of interest (ROI) contained in the forms. The drawback of this method is that it only recognizes the machine-printed tax forms and the performance is decreased if the keyword of interest is handwritten.

By finding pairs of corresponding form-lines on a reference form and a filled form, Bohnacker et. al described a tree search scheme to determine the location of form fields [1]. Similarly, Hirano et. al proposed a framework to extract fields of interest from faxed forms with distortions [2]. In this framework, the lines were extracted initially and different type of key points, such as corners, start points and end points were located according to the position of lines. Then the affine transformation was applied for image registration and the field was extracted on the transformed image. The potential problem of this method is that the registration of images suffers from falsely detected key points that are introduced by character strokes. Also, the affine transform often fails to capture the distortion of forms in real applications.

Inspired by Scale-invariant feature transform (SIFT) which is widely used in the field of computer vision and recognition, such as face recognition and image registration, we present a two-step form registration and field of interest extraction method in this paper which can effectively detect the location of data entries in the mix-type historical forms.

The rest of the paper is structured as follows. In Section 2 we describe the properties of our corpus. Section 3, describes the details of Affine SIFT (ASIFT), followed by the description of the proposed framework of data extraction in section 4. The experimental results and conclusions of our work are covered in section 5 and , respectively.

[1]The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

[2]Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

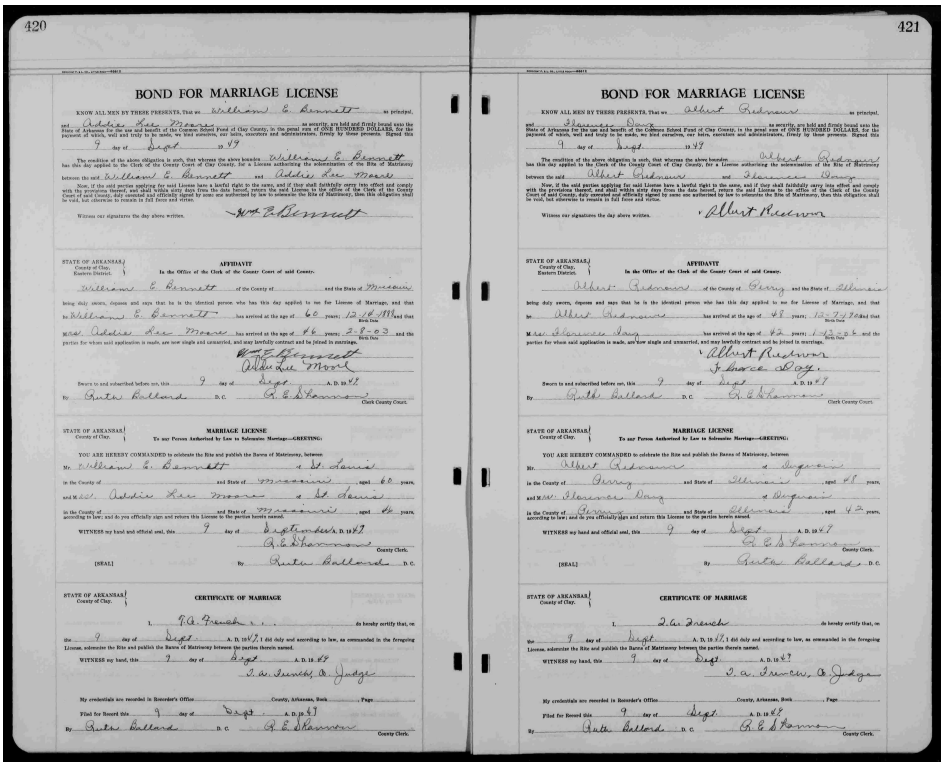


Figure 1: An example of mixed-type historical form in our corpus.

2. CORPUS PROPERTIES

For our experiments, we used 7000 historical Arkansas marriage record forms which are scanned images with resolution of 300 dpi and contain both machine-printed and handwritten text within the form images. Besides the blur and noise caused by down sampling during scanning, the images in our corpus suffer from other degradations, such as distortion, rotation, partially occlusion, etc. One of the example images from our corpus is shown in Fig. 1 which is a typical form with two identical pages in the same image. From this figure we can see that all fields of interest in the form are handwritten surrounded by machine-printed text.

We found that although the printed content of each form is similar, the layout of the images in the data set are not identical. Thus, prior to image registration and data extraction, we manually selected a total of 25 forms with different layouts and manually labelled the regions of interest. In our experiments, the data of interest are family tree related information, such as the bride and bridegroom's name, the resident of bride and groom, marriage date, etc. Table 1 lists all data entries we are interested in and in Fig. 3 we show the corresponding location of each entry in a template image using blue boxes.

3. KEY POINTS EXTRACTION AND REGISTRATION

Most image registration algorithms are based on extracting salient features from a document image. Our approach uses an affine scale-invariant feature transform (ASIFT) method to find salient points in forms. Generally, scale-invariant feature transform (SIFT) algorithm transforms an image from the original domain into a new domain which is represented by a set of feature vectors. These features are invariant to image translation, zoom, rotation, illumination, etc. and are robust to local geometric distortion. Thus, SIFT based methods are widely used in the computer vision areas such as face recognition and object tracking [3].

As one of advanced versions of SIFT, Affine-SIFT method introduces two more parameters for the feature descriptor to represents the angles of the camera axis orientation. [4]. By using these two new parameters, the ASIFT method can easily identify features even under large affine distortions which effectively increases the number of key points extracted for a given reference image and inquiry image.

Considering the merits of ASIFT and the distortion of form images we may encounter for the data set, we use ASIFT based method to extract key points for standard template images and store those key points as references. For a test image, we first calculate the height and width ratio of the image to determine whether the form is double-page or single-page. For a double-page form image, we estimate the binding line of each page according to the vertical profile of the form image and divide the image into two single pages. To each single page, a set of key points are extracted using ASIFT method and compared to the reference key points of each template. The key point matching is carried out by using nearest neighbor searching with Euclidean distance. Those matched points whose distance ratio is greater than a predefined threshold are rejected.

Entry Name	Description
BR_NAME	Bride's name
GR_NAME	Groom's name
BR_STATE	The state of bride
BR_CITY	The city of bride
GR_STATE	The state of groom
GR_CITY	The city of groom
YEAR	The year of marriage
MONTH	The month of marriage
DAY	The day of marriage

Table 1: Records extracted from corpus.

For each test image, the top-1 template which has the maximum number of match points with the test image is chosen as the candidate template. If the number of matching points between test image and the candidate template is smaller than a threshold (10 in our experiment), the test image is considered as an unknown form and human interaction is needed. Based on those matched points between candidate template and test image, we apply a perspective projection on test image which is defined by Eq. 1:

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & h_{1,4} \\ h_{2,1} & h_{2,2} & h_{2,3} & h_{2,4} \\ h_{3,1} & h_{3,2} & h_{3,3} & h_{3,4} \\ h_{4,1} & h_{4,2} & h_{4,3} & h_{4,4} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = H \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (1)$$

where $\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix}$ is the projected point, $\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$ is the original point in homogeneous coordinates system, respectively, and matrix H is perspective transform matrix or homograph matrix which can be estimated according to solution of linear least squares of Eq. 2:

$$\lambda_i x_i' = H x_i \quad (2)$$

where λ_i is the scale factor in homogeneous coordinates system, x_i' and x_i are matched points pair between template and test image. More details of calculating homograph matrix can be found in [7].

4. DATA EXTRACTION

The regions of interest (ROI) of the test image after perspective projection are very close to the ROI of template even if the original test image has severe affine distortion. To ensure the final ROI are extracted precisely, the second round of image registration in local area is implemented before the final data extraction.

At this step, we can determine the location of each ROI on the test image based on the corresponding ROI on the template image. Next, key points within an enlarged area around ROI for both test image and template image are extracted by using ASIFT method again. The key point matching and perspective projection procedures described in section 3 are now applied to the ROI. The final entries

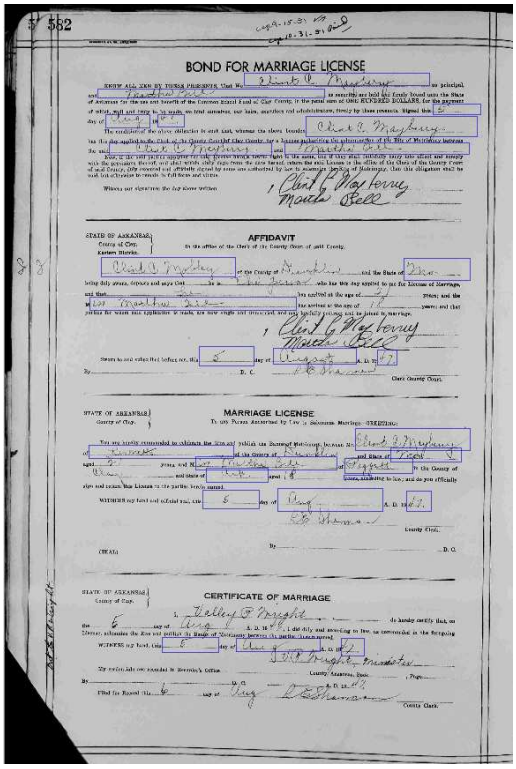


Figure 2: A template image with the corresponding ROI.

of interest can be easily located after this two-pass image registration.

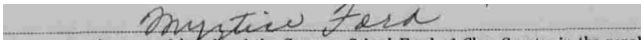
5. EXPERIMENTAL RESULTS

In our experiment, we used the proposed method to extract fields of interest from Arkansas marriage data set. The procedure can be briefly described as:

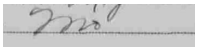
1. Select 25 template images cover the different layouts through the data set;
2. Extract key points from template images as references;
3. Extract key points from test image and find the best matching template;
4. Implement perspective projection for test image on page level;
5. Initially determine the location of each ROI of test image;
6. Extract key points from local area around each ROI;
7. Implement perspective projection for each ROI and determine the final location of each field of interest.

In Fig. 3, we illustrate the preliminary results of sample data which are extracted using the proposed method.

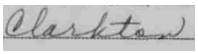
From these sub-figures we can see that the fields of interest within the form are correctly extracted and can be used for



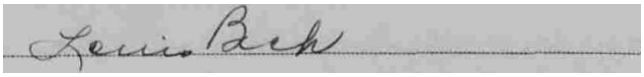
(a) Bride's name: Myrtice Ford



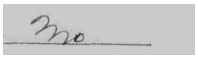
(b) Bride's state: Missouri



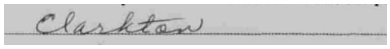
(c) Bride's city: Clarkton



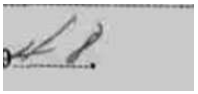
(d) Groom's name: Louis Beck



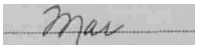
(e) Groom's state: Missouri



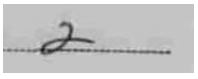
(f) Groom's city: Clarkton



(g) Marriage year: 1948



(h) Marriage month: March



(i) Marriage day: 2

Figure 3: The results of extracted data.

downstream processing such as optical character recognition (OCR) and information retrieval (IR).

6. CONCLUSIONS

In this paper, we presented an effective method for extracting entries of interest from mixed-type historical forms. The two-pass image registration scheme to determine the location of each data field of interest on the image provides reliable performance for key points extraction and image registration. Future work will focus on improving OCR for robust recognition of content in such images.

7. REFERENCES

[1] U. Bohnacker, J. Schacht, and T. Yucel. Matching form lines based on a heuristic search. In *Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on*, volume 1, pages 86 –90, aug 1997.

[2] T. Hirano, Y. Okada, and F. Yoda. Field extraction method from existing forms transmitted by facsimile. In *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, pages 738 –742, 2001.

[3] D. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150 –1157, 1999.

[4] J. Morel and G.Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2, 2009.

[5] H. Sako, M. Seki, N. Furukawa, H. Ikeda, and A. Imaizumi. Form reading based on form-type identification and form-data recognition. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pages 926 – 930, aug. 2003.

[6] J. Xu, V. Singh, V. Govindaraju, and D. Neogi. A hierarchical classification model for document categorization. In *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, pages 486 –490, july 2009.

[7] Z. Zhang. Estimating projective transformation matrix. Microsoft Research Techical Report MSR-TR-2010-63, 2010.