

A Hierarchical Method for Automated Identification and Segmentation of Forms

S. Mandal, S. P. Chowdhury, A. K. Das
CST Department

B. E. & Sc. University, Shibpur
Howrah 711 103, INDIA

{sekhar, shyama, amit}@becs.ac.in

Bhabatosh Chanda
ECS Unit

Indian Statistical Institute
Calcutta 700 035, INDIA

chanda@isical.ac.in

Abstract

In this paper we propose a fully automatic hierarchical method for identification of forms using global as well as local features. Moments of certain orders are considered as global shape features and are utilised to reduce the search space by selecting a subset of forms present in the database. The type of the candidate form is then identified within this subset through detail analysis using local geometrical and topological features. The candidate form is then segmented to extract the user-filled information.

1. Introduction

Though we are approaching an era of paperless office where every document will be electronically generated and processed; till date machine- or hand-printed form demands time consuming and costly processing. This is particularly true for a populous country and commonly used forms like tax return and application forms which are used in millions. Thus, automatic processing of forms is essential to reduce the time and cost involved. Moreover, storage, searching, browsing, distribution and dissemination of information provided via the paper form can be done more efficiently.

In any form processing system, that handles a large variety of forms, the objective is to identify the type the candidate form belongs to and to extract user-filled data from that. A blank form (i.e., form without any user-filled data) and a set of filled forms of same type may be used to determine reference prototype. These reference prototypes are stored in a prototype database and are compared with the instances of a filled-in input form. Storing the reference prototypes in the database and comparing them with the candidate form demands a suitable representation.

In this paper we propose an automatic hierarchical form processing system which takes care of real-life constraints to identify the type of a form in real-time and to extract the user-filled data. Use of moments as shape features is well studied [4]. It is also known that the projection signature retains the shape information, which is conformed by the ex-

istence of image reconstruction algorithm from projection data [12]. Here also global shape features of the form is extracted using moments of its horizontal and vertical projection profiles. The search space for identification of the form is sharply reduced using these global features. Thereafter, detail layout structures that are already posted as a part of reference prototypes from blank forms are matched with that of the the user-filled form being processed. This two-stage hierarchical approach provides a high degree of performance without incurring much computation.

2. Past work

There are many propositions for form processing available in the literature [11, 8, 13, 14, 9, 6, 5, 2, 3]. In most of the cases features related to lines and boxes are extracted. These features could be length, width, positions, crossing types and counts. The problems with these approaches are that the feature vector does not reflect the hierarchical layout of the forms. Another type of representation tries to utilise hierarchical structure of blocks as an X-Y tree [2, 14, 7]. X-Y tree representation is possible only for forms with boxes and thus has limited application area as many forms have horizontal marker lines instead of boxes. We discuss below the scheme adopted by one representative of each of the two different approaches.

Duygulu et al. [2] proposes an X-Y tree based hierarchical representation of form containing only distinct rectangular blocks. The proposition is very effective for the forms which are designed to be processed in the computer as all the fields (i.e., user data area) are distinctly enclosed in boxes. All such algorithms [1, 2] are not applicable to the forms which are partly or wholly void of rectangular blocks. Moreover, two different forms with similar block hierarchy leads to ambiguity.

Fan and Chang [3] proposed an approach for form document identification using features based on horizontal, vertical relationship matrices and crossing matrix to create prototype. The approach can be theoretically be applied to forms with boxes and horizontal markers. However, the al-

gorithm is very much dependent upon the appropriate selection of the threshold value which is not easy to set in practice. Moreover dotted lines are not considered which is again a restriction for real life applications. We have modified the relationship matrices proposed in [3] so as to cope up with real life situation and extended the process further to include segmentation after identification of the form type as discussed in the next section.

3. Proposed work

The major problem of the form identification is high computation cost and low performance without OCR. This is mainly because of detail structural (and sometimes semantical too) analysis required to identify the given form with the stored prototypes. Computation naturally increases with the number of prototypes stored in the database. Secondly, as the number of prototypes increases, their distinctiveness decreases which leads to more ambiguity and, hence, more errors.

We propose a hierarchical method for identification of forms which alleviates these problems to a large extent. First, a small subset of available forms is selected from the database in a way to guarantee the presence of the type of the form being processed in the subset. This is done by exploiting the moment-based global shape features. Detail analysis is performed next with forms of this subset, using the line (vertical and horizontal) structures and crossing types to identify the exact type. After identification of the type of the form, we segment the user-filled information from the candidate form by consulting the structure and pre-printed data of the identified form. It may be noted that any form has two parts: Pre-printed field names, and space to provide user-specific information. Preprinted text, other than the field names, may also be present as instructions to fill in the form or as clarification to some of the fields. These pre-printed data should be extracted and kept in the prototype database to help form segmentation.

3.1. Form layout: Observations

Forms are available in numerous styles. We have examined many commonly used forms including applications of different types, railway reservation, tax return etc. and we have made the some observations regarding the form layout.

The observations have led us to classify the forms into two primary types; F1 and F2 (see fig. 1(a) and (b)). F1 type has only boxes to contain the characters. F2 has horizontal lines as markers above which the information is filled-in. However, in real life, forms show a mixed structural property where a portion may be considered as F1 and the other could be F2 (see fig. 2(a)). It may be noted that there are

some forms where no boxes or lines are present. In such forms only the field names are printed on the left and space for user-specific information on the right. This type of form is not considered in this work.

3.2. The system

A form processing system capable of handling large varieties of forms has two phases: (i) Generation/upgradation prototype database, and (ii) identification and segmentation of candidate form. Two types of features are extracted from each form either for inclusion in the database or for identification of the type of form. These are (i) moment based global shape features, and (ii) structural shape features.

3.3. Global shape feature

For extraction of the global shape features a set of filled forms of each type is used. The 2nd and 4th order moments of the vertical and horizontal projection profiles of the form images are taken as global shape features. Shape features, in general, should be invariant to scale, rotation, and translation. Fortunately, for form processing, scaling problem does not arise. In practice skew is also not a significant problem. So, we have to take care of translation problem only.

Three sub-steps of global shape feature computation are elaborated below.

i) Selection of training samples: For a particular type of form we take n number of user-filled instances. We have used $n = 15$. It may be noted that samples have small skew and are filled by different users.

ii) Computing average horizontal and vertical projections: For i -th type of form, at first, all the horizontal projections $h_k^{(i)}$ (for $k = 1, \dots, n$) are registered. We have taken the first sample $h_1^{(i)}$ as the anchor and calculated the correlation of others with respect to this as:

$$\rho_{1k}(x) = \sum_{r=0}^{R-1} h_1^{(i)}(r) \times h_k^{(i)}(r+x)$$

where R is the number of rows in the anchor image. Let

$$\tilde{x}_k = \text{arg} \left[\max_x \{ \rho_{1k}(x) \} \right]$$

Thus, \tilde{x}_k is the amount of shift required by horizontal projection of the k -th image to be aligned with that of the first one. Hence, average horizontal projection is obtained as

$$\bar{h}^{(i)}(x) = \frac{1}{n} \sum_{k=1}^n \tilde{h}_k^{(i)}(x + \tilde{x}_k)$$

Exactly similar procedure is followed to compute average vertical projection $\bar{v}^{(i)}(y)$.

014753

BENGAL ENGINEERING COLLEGE
(A DEEMED UNIVERSITY)
HOWRAH - 711103, INDIA

APPLICATION FOR ADMISSION TO
POSTGRADUATE COURSES IN
ENGINEERING DURING THE
ACADEMIC SESSION 2004-05

Please put tick mark (✓) in appropriate single boxes

1. Name of the Candidate in full in block letters
 Shri Shrimati
SUVANJAYAR
 First Name
DAS GUPTA
 Middle Name
 Surname

2. Category of Candidate
 General SC ST

3. Date of Birth
 DD MM YYYY
28 08 1981

4. Names of the Qualifying Degree and the Discipline
B.E. (Information Technology)

5. Status of the Candidate in respect to the Qualifying Degree
 Passed in the year _____
 Shall appear in the year 2004
 Appeared in the final Examination Result expected in the year _____

6. Branch in which admission is sought
Computer Science & Technology

7. Performance in GATE Registration No. 1741204
 Year of Examination 2004
 Score (Percentage) 97.00

8. Other branches in which the Candidate has applied for admission in the year 2004
 1. _____
 2. _____

9. Area of Specialization sought after by the candidate (in order of preference)
 1. Software Technology
 2. and Computer Engineering

10. Academic Record (From 10th to the Qualifying Examinations)

College / Institute	University / Board / Examining Authority	Year of Passing	Division / Class	Marks / % Grade
Kendrapada Vidya Mandir, Kendrapada	WBSE, H.S.S.	Upto 10th sem	1st	82.4%
Siliguri Institute of Technology	University of North Bengal	2004	1st	84%

11. Whether the Candidate is sponsored by the Institute or not?
 Yes No
 If yes, please give details of industrial experience in a separate sheet.

Parent's / Guardian's Name: Sri. Das Gupta
 Relationship with the Candidate: Father
 Address for Communication: 10, Anandapur, Udayagram
P.O. - Dhanbad, Dist - Hooghly
 Pin 712124 Phone 963031493

I do hereby declare that:
 (a) The information given above is true to the best of my knowledge and belief;
 (b) I have not been admitted to the Postgraduate course of any other University/Institute on the basis of the GATE score under release;
 (c) I undertake to abide by the Rules and Regulations of the University, if selected for admission.

(a)

014753

BENGAL ENGINEERING COLLEGE
(A DEEMED UNIVERSITY)
HOWRAH - 711103, INDIA

APPLICATION FOR ADMISSION TO
POSTGRADUATE COURSES IN
ENGINEERING DURING THE
ACADEMIC SESSION 2004-05

Please put tick mark (✓) in appropriate single boxes

1. Name of the applicant: Sankar Mandal
 2. Surname at birth (if any): Mandal
 3. Other names (where applicable):
 4. Date of birth: 02-12-1966
 5. Place of birth (town, city & country): 66, 1, Pexu, Eastson

2. Father's / Mother's / Spouse's name & Nationality:
 a. Father: Sankar Mandal
 b. Mother: Savitri Mandal
 c. Spouse: Mandira Mandal
 (where applicable)

3. Address:
 a. Present: D-227, B.E. College Campus, P.O. Behura, Gopabandhu, Howrah - 711103
 b. Permanent: DS
 c. Address in India (verifiable):

Paste your PHOTO here

(Please Type or Print)
 (Please read the instructions carefully before filling the application)

1. Full Name: Sankar Mandal Mukherjee
 (First) (Last)

2. Last name at birth (if different): _____

3. Marital Status: Married/Unmarried Unmarried

4. If married give maiden name: _____

5. Date of Birth: 02 / 12 / 66 Sex: Male

6. Sex: Male/Female

7. Place of birth (City, State & Country): Kolkata, West Bengal, India

8. Current Nationality: Indian

9. Are you a permanent/long-term resident in this country? Yes No Yes
 If yes, please furnish photocopy of such document (GREEN-CARD/Long-Term Visa status):
 (For Non-L/S passport holders only)

10. Nationality at birth: Indian 11. Any other nationality held at present: _____

12. Present Address: 66, 1, Mukherjee Road, Kolkata, West Bengal, India

13. Phone: 91 94 15 46 51 2 (Home) 2 67 8 9 5 5 7 (Work) 2 5 6 4 1 1 2 2

(a)

EMBASSY OF INDIA
CONSULATE GENERAL OF INDIA
HIGH COMMISSION OF INDIA
Consular Wing

APPLICATION FORM FOR THE GRANT/RENEWAL OF PIO CARD
(TO BE FILLED IN DUPLICATE)

Please read the instructions including the eligibility criteria carefully before filling the form. The Embassy/Consulate is not responsible for any error, which may lead to the non-acceptance of the application.

1. a. Name of the applicant: Sankar Mandal
 b. Surname at birth (if any): Mandal
 c. Other names (where applicable):
 d. Date of birth: 02-12-1966
 e. Place of birth (town, city & country): 66, 1, Pexu, Eastson

2. Father's / Mother's / Spouse's name & Nationality:
 a. Father: Sankar Mandal
 b. Mother: Savitri Mandal
 c. Spouse: Mandira Mandal
 (where applicable)

3. Address:
 a. Present: D-227, B.E. College Campus, P.O. Behura, Gopabandhu, Howrah - 711103
 b. Permanent: DS
 c. Address in India (verifiable):

Paste your PHOTO here

(b)

EMBASSY OF INDIA
CONSULATE GENERAL OF INDIA
HIGH COMMISSION OF INDIA
Consular Wing

VISA APPLICATION FORM

For Office use only

Paste your PHOTO here

(Please Type or Print)
 (Please read the instructions carefully before filling the application)

1. Full Name: Sankar Mandal Mukherjee
 (First) (Last)

2. Last name at birth (if different): _____

3. Marital Status: Married/Unmarried Unmarried

4. If married give maiden name: _____

5. Date of Birth: 02 / 12 / 66 Sex: Male

6. Sex: Male/Female

7. Place of birth (City, State & Country): Kolkata, West Bengal, India

8. Current Nationality: Indian

9. Are you a permanent/long-term resident in this country? Yes No Yes
 If yes, please furnish photocopy of such document (GREEN-CARD/Long-Term Visa status):
 (For Non-L/S passport holders only)

10. Nationality at birth: Indian 11. Any other nationality held at present: _____

12. Present Address: 66, 1, Mukherjee Road, Kolkata, West Bengal, India

13. Phone: 91 94 15 46 51 2 (Home) 2 67 8 9 5 5 7 (Work) 2 5 6 4 1 1 2 2

(b)

Figure 1. Example of forms; (a) Boxes only (F1 Type) to contain user-filled data and (b) Only lines (F2 type).

Figure 2. Example of forms; (a) Boxes and lines mixed; (b) F2 type and very similar to the form shown in fig. 1(b).

iii) Computing second and fourth order central moments:

Second and fourth order moments of $\bar{h}^{(i)}(x)$ and $\bar{v}^{(i)}(y)$ are computed and stored in the prototype database. Let us denote them by $m_h^{(i)}(2)$, $m_h^{(i)}(4)$, $m_v^{(i)}(2)$ and $m_v^{(i)}(4)$

3.4. Structural features

Structural features are extracted from blank form of each type using the following operations.

Step 1: Pre-processing

It consists of subtasks like, (i) noise removal and skew correction, (ii) converting all dotted lines to solid lines using morphological closing, and (iii) extending broken lines of the boxes to meet with the nearest line orthogonal to it. Then all the horizontal and vertical line segments are extracted using morphological sieving. It may be noted that there may be adjacent boxes along rows or columns or both. As a result horizontal and vertical lines may form 9 different types of crossings [3] as shown in figure 3.

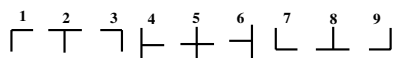


Figure 3. Nine types of crossings.

Step 2: Formation of line and crossing matrices

Next step is the extraction of positions of lines and boxes to get the layout features. The detail information on relative positions of horizontal and vertical lines are extracted and stored in double valued horizontal and vertical relationship matrices (DVHRM and DVVRM). Crossing relationship matrix (CRM) is generated based on the crossing type and position. It may be noted that this was originally proposed in [3] which is modified here to make it robust to geometric transformation and noise. These matrices represent characteristics of the form layout. Hence, these are used in detail analysis.

The prototype database should also keep the pre-printed text (field names etc.) information for extraction of user-filled portions and is elaborated next.

3.5. Pre-printed text labelling

In this step pre-printed text are extracted and are stored along with their positions. These pre-printed text, especially the field names, along with the blank boxes or the horizontal line markers against which the field names are printed constitute what we call 'meta-form'. These blank meta-forms are stored in the form database.

3.6. Form identification

This is done by comparing features of candidate form with that of prototypes already stored in the database hierarchically. As pointed out earlier, moment-based global shape features are first compared with the database to get a small subset of forms. We select S prototypes out of N prototypes stored in the database ($S \ll N$) corresponding to lowest values of distances defined as

$$d(F, F_i) = \sum_{j=2,4} \frac{|m_h^{(i)}(j) - m_h(j)|}{|m_h^{(i)}(j)| + |m_h(j)|} + \sum_{j=2,4} \frac{|m_v^{(i)}(j) - m_v(j)|}{|m_v^{(i)}(j)| + |m_v(j)|} \leq t$$

Where F and F_i are candidate form and the i -th prototype, respectively and t is acceptable tolerance in dissimilarity among the similar forms. This dissimilarity may be introduced due to printing, digitization etc. If the dissimilarity $d(F, F_i)$ is greater than t for all i , then the candidate F form is considered as unknown and may be included as a new prototype. The value of t is set through rigorous statistical study of large number of similar forms of each type. In the next step, the structural feature matrices of the candidate form are compared with that of the members of the small subset already available from the first stage, to arrive at the final decision on the form type.

3.7. Form segmentation

In this step we extract the user-filled information and put them in the copy of the corresponding meta-form stored in the Form database. The details of segmentation is given below.

- All the boxes are examined using the raster scan direction.
- User-filled characters and words are extracted from the boxes.
- For each horizontal line in the blank form
 - Fit the largest rectangular window within the blank space above the horizontal line, and
 - get the user filled information whatsoever available within this window.

4. Results and discussion

We have trained our system for 40 different types of forms. A couple of forms are similar in look (see fig. 1(b) and 2(b)). For each type, one blank form and 15 filled-in forms are used to generate the prototype. The system is tested with 10 filled-in forms of each type i.e., 400 forms in total which are not used in training. As stated earlier a hierarchical identification is followed. At first moment based features returns a subset of S possible types, according to the S lowest values of dissimilarity (as defined in sub-section 3.5). Then DVHRM and DVVRM of the candidate form are

No of forms	1st	2nd	3rd	4th	5th
400	287	73	24	9	4

Table 1. Count of ordinality number of the candidate form within the subset.

No of forms	1st	2nd	3rd	4th	5th
400	301	33	17	6	8

Table 2. Count of ordinality number of the candidate form using only the local features.

compared with that of the S possible types already available from the first stage. The CRM is also compared in a similar manner. The technique for comparison of DVHRM, DVVRM and CRM matrices of the query form and database forms is explained in [10]. The strength of the hierarchical method can be verified from Table 1. The table shows the frequency of occurrence of the ordinality number of the candidate form in the selected subset. Hence, it shows the possibility that the global feature (moments) used to select the subset of forms almost always contain the type of form being processed. Out of 400 form only 3 forms or 0.75% are not present in the subset of first 5 selected types as shown in the table. This error is due to the forms that are filled partially by the user for some reason or other. As the candidate form is almost surely mapped within the first 5 possible types we have taken $S = 5$ to reduce the amount of computation in the next stage.

In the next stage, utilising the local features, detail analysis is done to find out the exact type of the candidate form from the subset selected in the first stage. The performance of this two stage hierarchical approach is very encouraging as we are able to detect the exact type of 387 (i.e. 97%) forms out of 400 input forms. After identification we have extracted the user filled information using procedure mentioned earlier.

Since the detail analysis utilising the local features is used to pinpoint the exact type, it may be argued to use this stage only bypassing the use of global features. However, Table 2 shows the result of the frequency of occurrence of the ordinality number of the candidate form utilising only the local features. The problem utilising the local features is evident from the table and we see that the local feature method leads to more computation and more ambiguity.

5. Conclusion

We have presented a fully automatic hierarchical form processing system. The proposed method is based on rep-

resentation of the form in terms of global as well as detail structural features and it is applicable to most of the commonly used forms. However, the proposed method will fail to discriminate two semantically different forms having same layout, as the algorithm does not take any kind of help from OCR. It may also be mentioned that we did not try to incorporate procedure to alleviate form dropout problem and this may be added in future.

References

- [1] F. Cesarini, M. Gori, S. Marini, and G. Soda. Structured document segmentation and representation by the modified x-y tree. In *Proc. of 5th ICDAR*, pages 563–566, 1999.
- [2] P. Duygulu and V. Atlay. A hierarchical representation of form documents for identification and retrieval. In *SPIE, Electronic Imaging 2000, Document Recognition and Retrieval VII*, San Jose, USA, January, 2000.
- [3] K.-C. Fan and M.-L. Chang. Form document identification using line structured based features. In *14th Int. Conf. on Pattern Recognition*, Brisbane, Australia, August, 1998.
- [4] R. C. Gonzalez and R. Wood. *Digital Image Processing*. Addison-Wesley, Reading, Mass., 1992.
- [5] P. Heroux and S. Doermann. Classification method study for automatic form class identification. In *14th Int. Conf. on Pattern Recognition*, Brisbane, Australia, August, 1998.
- [6] O. Hori and D. Doermann. Robust table-form structure analysis based on box-driven reasoning. In *ICDAR95*, pages 219–221, 1995.
- [7] Y. Ishitani. Flexible and robust model matching based on association graph for form image understanding. *Pattern Analysis and Application*, 3:104–119, 2000.
- [8] J. Lin, C. Lee, and Z. Chen. Identification of business forms using relationships between adjacent frames. *Machine vision and Applications*, 9:56–64, 1996.
- [9] J. Liu and X. Wu. Description and recognition of form and automated form data entry. In *Proc. Third Int. Conf. on Document Analysis and Recognition, ICDAR'95*, pages 579–582, 1995.
- [10] S. Mandal, S. P. Chowdhury, A. K. Das, and B. Chanda. Automated detection and segmentation of form document. In *Proc. of 5th International Conference on Advances in Pattern Recognition; ICAPR-2003*, pages 284–288, Dec. 10–13, Calcutta, India, 2003.
- [11] J. Mao, M. Abayan, and K. Mohiuddin. A model based form processing sub-system. In *Proc. of 13th Int. Conf. on Pattern Recognition, ICPR'96*, pages 691–695, Vienna, Austria, August '96, 1996.
- [12] A. Rosenfeld and A. C. Kak. *Digital Picture Processing*, volume Vol II. Academic Press, N.Y., 1982.
- [13] S. Shimotsuji and M. Asano. Form identification based on cell structure. In *Proc. 13th Int. Conf. On Pattern Recognition*, pages 793–795, Vienna, Austria, August, 1996.
- [14] T. Watanabe, Q. L. Luo, and N. Sugie. Layout recognition of multi-kinds of table-form documents, 17. *IEEE transactions on Pattern Analysis and Machine Intelligence*, (4):432–446, 1995.