Form Document Representation And Identification *

Pınar Duygulu *, **Volkan Atalay** * , **Ebru Dincel** *+

* Dep.of Computer Eng., Middle East Technical University, Ankara, TR-06531 Turkey
+ TÜBİTAK-BİLTEN, Middle East Technical University, Ankara, TR-06531 Turkey
e-mail : {duygulu,volkan,dincel}@ceng.metu.edu.tr

A form processing system aims to extract meaningful data from a form document for office automation [1, 2, 3]. The main interest is to extract user filled-in data which is considered to be meaningful. However, in order to perform such a task, the structure of the form should be known in advance. The form structure can be obtained by processing a blank model form on which no user filled-in data exists. The model form may be stored in a database and when an instance form is presented to the system, its type can be identified by matching with one of the model forms in the database. Meaningful data can then be automatically extracted. A form document can be easily represented by its image. But, such a physical information may not be appropriate, since the form or its image risks to be modified geometrically or to be distorted due to printing or digitization. On the other hand, logical structure represents the semantics of the form and same logical structure can be formatted in a variety of physical layouts. The geometrical structure should be mapped to a logical structure by considering the logical relations.

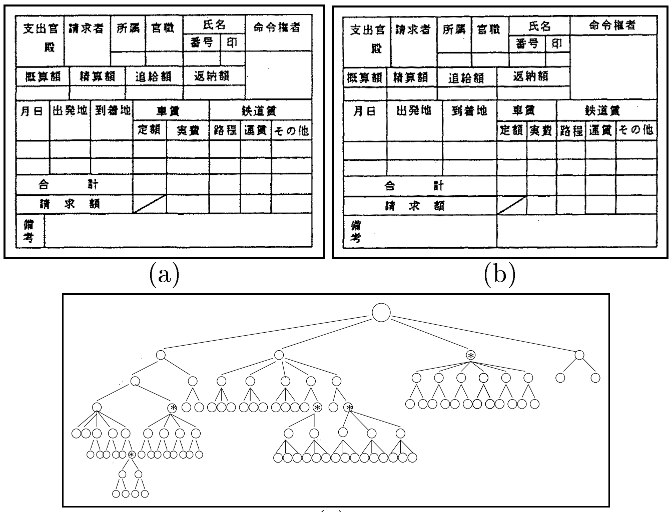
In this paper, our aim is to develop a logical representation for form documents and to use the representation for identification. We propose a heuristic algorithm in order to transform geometric structure of a form into a logical structure by using horizontal and vertical lines which exist on the form. The logical structure is represented by a hierarchical tree. The hierarchy of the tree corresponds to the hierarchy of the blocks in the form document. The proposed representation is close to human point of view for the form structure. Logically same forms have the same hierarchical tree structure. Also, geometrical modifications and slight variations on a form are handled by the proposed representation.

A hierarchical structure is proposed to represent the logical layout of a form by using horizontal and vertical lines. The main aim is to partition the form into blocks which can further be divided until cells (the smallest blocks) are reached. The partitioning results in a tree where the root is the form itself and the leaf nodes correspond to the cells. The heuristic behind the approach is that the blocks which contain similar information are grouped together and these group of blocks are separated by the lines relatively longer than the others. Such lines are defined as baselines. Thus, the information about how to partition a block is given by the baselines. However, not all baselines provide this information, but the orthogonal lines give clues about which baselines are separators. In

* This work is partially supported by TÜBİTAK-BİLTEN (Scientific and Technical Research Council of Turkey - Institute of Information Technologies and Electronics) under the project 96-14-030.

order to achieve a sequence of block partitioning, a switching of horizontal and vertical divisions is proposed.

The approach is top-down and no domain knowledge such as the preprinted data or filled-in data is used. At the end, two forms which have different physical structures but the same logical structure will be associated to the same representation. An instance form is registered with one of the model forms in the form database if their hierarchical trees are identical. We give an example form document to clarify the proposed algorithm which is adapted from Watanabe *et al.* [4]. In fact, there are two form documents which are logically same, but physically different. When the proposed heuristic algorithm is applied to these form documents, the same tree is obtained (c.f. Figure 1).



(a) (b)

(c)

Figure 1. Multi-kinds of the same table form document (adapted from Watanabe *et al.* [4]) which are represented by the same hierarchical tree.

References

1. Y.Tang, C.Yan, M.Cheriet, C.Suen, "Automatic Analysis and Understanding of Documents," in *Handbook of Pattern Recognition and Computer Vision*, pp. 625-654, 1993.
2. D.Wang, S.Srihari, "Analysis of Form Images," *First Int'l Conf. Document Analysis and Recognition*, Saint-Malo, France, pp. 181-191, Sept. 1991.
3. R.Casey, D.Ferguson, K.Mohiuddin, E.Walach, "Intelligent Form Processing System," *Machine Vision and Applications*, vol. 5, pp. 143-155, 1992.
4. T.Watanabe, Q.Luo, N.Sugie, "Layout Recognition of Multi-Kinds of Table-Form Documents," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.17, no.4, pp.432-445, April 1995.
5. P.Duygulu, V.Atalay, E.Dincel "A Heuristic Algorithm For Hierarchical Representation Of Form Documents, " accepted to be presented at *14th Int'l Conf. Pattern Recognition*, Brisbane, Australia, August 1998.