Perceptual Organization in Semantic Role Labeling

Prateek Sarkar Eric Saund

Perceptual Document Analysis
Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304
{psarkar,saund}@parc.com

September 30, 2005

Abstract

Documents are produced for the purpose of human interpretation. Human perceptual factors have played an important role in the design of documents — from the development of glyphs and scripts to the layout of visual components. OCR technology allows recovery of textual content from images of text but does not recover visual imformation encoded in layout. We explore the role of perceptual organization in the interpretation of documents and related computational challenges. A useful application area is semantic role labeling: the problem of assigning semantic roles to visual or textual structures in documents. We present a generic A^* search formulation that has been applied to reduce search times by orders of magnitude in a semantic role labeling problem.

1 What is Semantic Role Labeling

Documents, as records of human communication, are built up of visual components that play distinct roles in expressing the meaning of a document. While these roles are various, and visual artists are still discovering ways of communicating through images, major categories of semantic roles can be enumerated in the context of document processing. For example, business invoices contain visual components that communicate identity (address block, company logo), object relations (item-price relations, signator-signature pairing), special attention (highlighting, circling). Parsing a document image into its semantically significant components is the problem of Semantic Role Labeling. We develop automatic and semi-automatic approaches to Semantic Role Labeling especially for large volume image understanding and indexing projects.

2 What is Perceptual Organization

The science of Perceptual Organization, pioneered by Wertheimer, Koffka, and Köhler, studies patterns and principles by which humans organize visual stimuli into perceptual forms in ways that are quite universal across both peoples and situations. This science advocates the Gestalt principles of perceptual grouping, and theories of figure-ground separation. The common Gestalt principles of grouping by proximity, similarity, continuity, closure, symmetry, surroundedness apply to both design and analysis of visual documents. In all scripts, proximity informs the grouping of glyphs into words. Curvilinear alignment, and feature similarity are vital in perceiving more complex textual blocks as visual objects rather than just a collection of glyphs. Enclosures and separators are frequently used to guide perceptual and semantic grouping.

3 Perceptual Organization in document image interpretation

The interpretation of complex document layout structure has remained a research subject. Unlike the parsing of one-dimensional text, image parsing even in very structured domains is algorithmically difficult, because of the lack of a natural ordering in

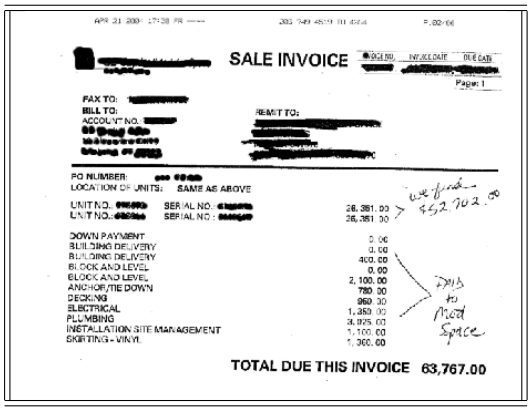

Figure 1: Perceptual queues in parsing a business invoice. Semantically significant information such as the logo and address areas, header alignment, tabular structure, presence of hand-written annotations, graphic separators are all visually salient features in the design of the invoice, and/or important cues in the analysis of such images.

two dimensions. Although very constrained grammar models (such as X-Y trees [NKSV93]) can keep parsing problem tractable, parsing of images in general presents an exponential search space. This exponential complexity derives from the exponential number of ways in which image primitives can group to form larger objects. Bottom-up grouping to date has been engineered for proximity, and special cases of curvilinear continuity (e.g., horizontal sequences of words form text lines). Non-overlapping boxes and convex hulls have also been used as criteria for pruning groupings in images [MV98].

If alternative groupings are to be pruned to keep the search tractable, we believe that such pruning should be heavily informed by perceptual organization principles, especially because document designs and conventions have evolved to suit human perception. Of course, perceptual grouping cues can also directly aid to resolve ambiguity in grouping. In other words they can form a vital component of the objective function rather than just pruning heuristics. Figures 2 show examples of images where layout of image primitives rely on perceptual grouping for their interpretation. Tree grammars are in general not sufficient for capturing the compositional structure of such images.

4 Research challenges

Hierarchical decomposition of document images (top-down or bottom-up) according to preset rules can help in analysis and sematic role labeling of a limited set of documents. Documents containing mainly textual material arranged in rectangular layouts have been the main targets of successful methods. In more complex documents (tables, maps, handwritten memos, drawings, photographic images) both segmentation, and the parsing of segments into meaningful structure is, in general, exponential in the number of alternatives. Interpretation such documents will require segmentation of images, and their interpretations to guide and inform each other. The principal challenges are:

- *A model of document structure:* The logical and visual structures of documents – constituents, correspondences and relationships – will have to be expressed in a language that allows for information from one domain to be extracted into the other.

- *Machine learnable, adaptive models for visual structure:* The challenge is to build models for the Gestalt grouping principles that can be repurposed for a variety of document image analysis applications, and use such models in the parsing of document images.

- *Smart search:* The interpretation of a given image is not the output of a number of pre-defined processes, but rather a search through the space of all possible segmentations, groupings and interpretations. This search process should be adaptive to the image at hand, generate and keep alive multiple plausible interpretations, incorporate a principled way of comparing alternative hypotheses, and be prudent regarding hypotheses that are necessary to evaluate (to beat the curse of exponential explosion).

5 Conquering the search problem

We shall provide a glimpse of the formulation of a specific semantic role labeling problem, and an A^* search algorithm that we have developed for its efficient solution. Consider an example where an image has been decomposed into its compositional atoms, *i.e.*, chunks of image that can take on a semantic label

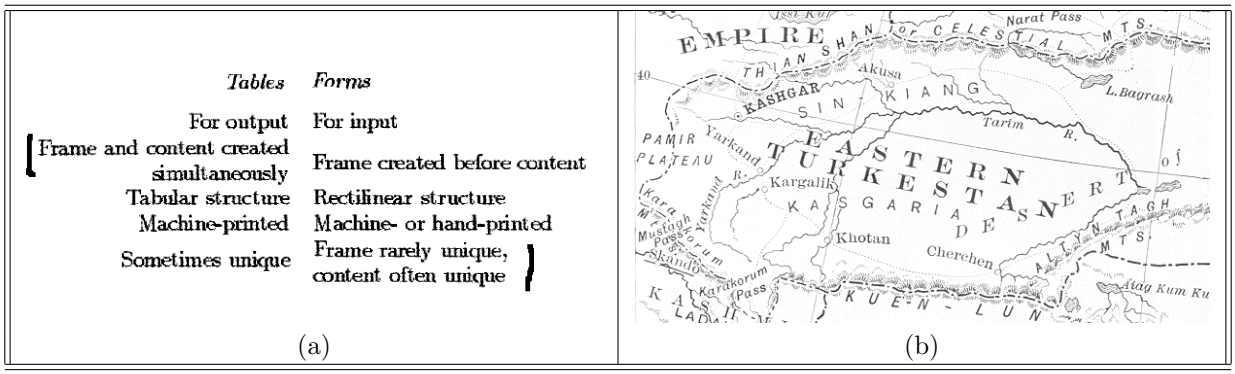

Figure 2: Examples of perceptual grouping in images. (a) The vertical alignment of text line groups is vital evidence that the marked text lines should be grouped into a single table cell. (b) The vertical alignment of text line groups is vital evidence that the marked text lines should be grouped into a single table cell. Parsing under such grouping cues is of exponential complexity.

without further subdivision. The problem is to assign labels $(y_1, y_2, \ldots, y_n)$ to a collection of observation atoms $(x_1, x_2, \ldots, x_n)$. In a typical pattern recognition formulation, the assignment is chosen by maximizing some objective function $f(x_1 \ldots x_n, y_1 \ldots y_n)$ over all possible labelings. If each atom can take on one of C label-values, the label assignment problem is a search through a the space of C^n label assignments. Restricting the nature of the objective function avoids the exponential explosion of the search space. For example, if the objective function is factorizable as the product of functions of the form $f_i(x_i, y_i)$, our problem reduces to labeling n observation independently, and the search complexity is nC. But often factorization of the objective function, if present, is not as simple. In addition to intrinsic features of x_i that influence the labeling y_i, we may have:

- *Factors coupling two or more labels* of the form $h(y_i, y_j, \ldots)$. Linguistic constraints in OCR are in this category. A Gestalt preference for repeated structure can be modeled with such such factors. In parsing a book page, at most one of all the text-word atoms may take on the label "page-number".

- *Factors coupling two or more observations* of the for $g(x_i, x_j, \ldots)$. These arise when the atomic observations have constraints on relative size, orientation, or color change (lighting variations). Style consistency [SN05, VN05] models require such factors to model shared fonts, color, or other characteristics of text.

- *More complex factors* such as direct dependence between labels and features of different atoms. For example, to label a dot in an image as a "bullet" induces preference for indentation on adjoining atoms, and alignment with other adjoining "bullet" atoms. Both indentation and alignment are features of groups of atoms.

Most perceptual organization principles induce such complex dependencies among atomic constituents of images. Various techniques have been developed in the graphical models literature for solving the joint labeling problem (e.g., loopy belief propagation, sampling, variational methods, and hybrid techniques.) However these techniques do not offer guarantees of joint optimality. We have developed an A^* search framework, where we find easily factorizable upperbounds for complex objective functions. In particular, when the upper-bound is expressed as a product of n functions $\bar{f}(x_i, y_i)$ the bound can be maximized in linear time. It is then used to guide the search in a best first way, thereby guaranteeing that the optimal solution will always be found. While the worst case complexity of this method is still exponential, the low average complexity can enable the solution of problems that are otherwise intractable.

6 Application: Indexing engineering drawings

A typical repository-indexing or metadata-extraction application requires scanned images of drawings to be indexed by a number of fields such as drawing number, drawing title, date of creation, author, company

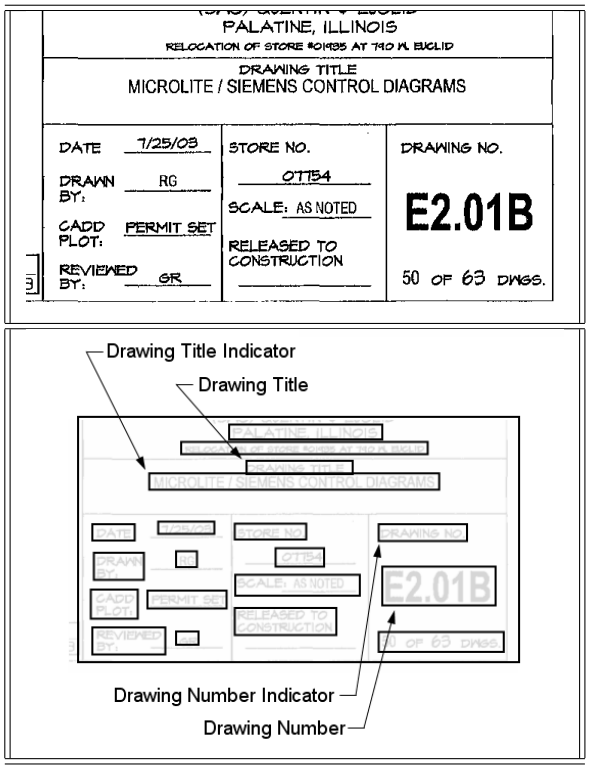

Figure 3: An example of a title-block image fragment showing the four relevant fields to be extracted, and the competing irrelevant text groups in the neighborhood.

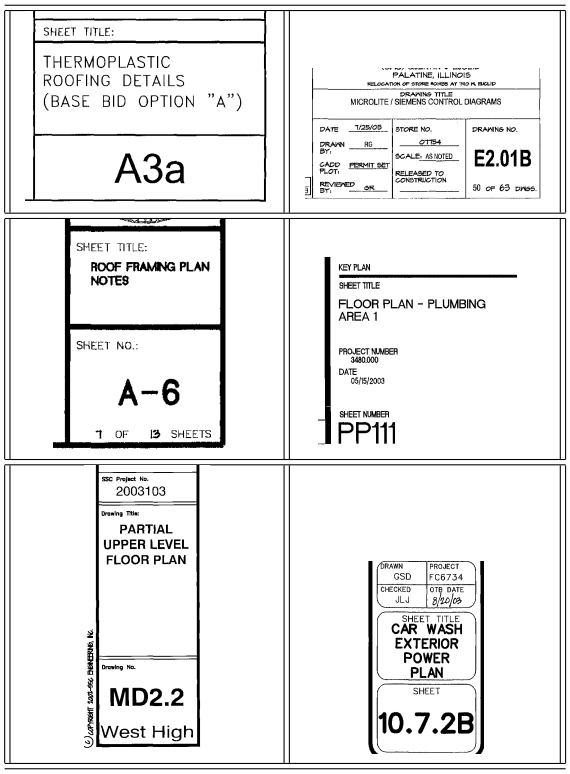

Figure 4: Examples of index fields in snippets of engineering drawings: drawing number and drawing title.

logo, site name, and scale. In our pilot technology development project we focused on extracting drawing numbers and drawing titles. Figure 3 shows an example of an image fragment that contains the drawing number and drawing title. While most drawings include drawing numbers and titles, they are printed at various locations, and in various formats and styles (Figure 4.)

Our objective is to locate these fields in the images without restricting input to a small number of layouts and styles. To this end, an input image is pre-processed to generate a number of candidates for the drawing number and title fields. The first step is perceptual grouping of foreground pixels into potential text fields. Connected components are first grouped to obtain word-like elements, which are further grouped into text-lines and multi-line text objects. Proximity, size similarity, alignment, and surroundedness (a text group does not straddle line enclosures) are used in the grouping process. In the typical engineering drawings in our dataset, this results in 30-200 text groups (see Figure 3.)Each resulting group in the image is sent to an external OCR engine, and any deciphered text is recorded.

We assume that each relevant field is grouped into a single unit, and not merged with other fields. In other words, each group found by our grouping algorithm, along with its OCR text, is treated as an atom for labeling. Our problem then is to label at most one of these groups as *drawing title (DT)*, and at most one as *drawing number (DN)*. To aid in the process, we also attempt to find a *drawing title indicator (DTI)* (*e.g.*, "Sheet Title",) and a *drawing number indicator (DNI)* (*e.g.*, "Sheet Number".) All remaining fields are to be labeled *irrelevant (IRR)*. We thus have $C = 5$ classes, $n \approx 50$ (on average) objects to label, and a search space of $\approx 50^5$ competing label-hypotheses.

A label-hypothesis is represented as $(l_{dn}, l_{dni}, l_{dt}, l_{dti})$ meaning that the group l_{dn} is labeled *DN*, group l_{dni} is labeled *DNI*, and so on. $l_{(.)}$ can take on values $1 \ldots n$ or 0, the latter indicating that the field is missing. Our problem is to find a label hypohesis that maximizes some score. The score (likelihood) of a label-hypothesis is assumed to

be of the following form:

$$L(l_{dn}, l_{dni}, l_{dt}, l_{dti}) \propto f_{dn}(l_{dn}) \cdot f_{dni}(l_{dni}) \cdot f_{dt}(l_{dt}) \cdot f_{dti}(l_{dti}) \cdot u(l_{dn}, l_{dni}) \cdot v(l_{dt}, l_{dti}) \cdot w(l_{dn}, l_{dt}) \quad (1)$$

The formula contains two kinds of functions:

- *Unary functions* f_{dn}, f_{dni}, f_{dt}, f_{dti}, which compute a goodness score (likelihood) that a group is a *DN*, *DNI*, *DT*, *DTI* respectively based on measurements that are properties of the group alone. Examples of such properties are *size*, *aspect ratio*, *position on page*, *regularity of size of group components*, *average stroke width*, *closeness of OCR text to expected textual content*, etc. In our pilot experiment we use only the last mentioned feature.

- *Pairwise functions* u, v, w, which compute scores on pairs of label candidates. These functions may thus depend on *relative location* and *relative alignment* of two groups, whether the two groups are *neighbors in a Delaunay triangulation*, whether they *share an enclosure* demarkated by line or white-space separators, etc. Pairwise functions can also be used to model mutual exclusivity constraints such as ($l_{dn} \neq l_{dt}$), *i.e.*, the same group cannot be both a drawing number and a drawing title. In our pilot experiments only the relative location of two labeled groups and mutual exclusivity were used in the pairwise functions.

In general, functions involving three or more groups can also be included. The overall score is proportional to the product of all such functions, and the winning label-hypothesis is the one that maximizes this product. If the objective function includes only unary functions, each component unary function can be maximized independently to arrive at the result. The pairwise (and higher order) terms, introduce complexity because such independent term-by-term maximization is no longer useful. Nevertheless, the pairwise functions are important for distinguishing, for example, the drawing number from the many other alpha numeric fields that appear on a page. So we follow the A^* solution suggested in the previous section, by replacing the pairwise terms by fixed upper bound to obtain an overall upperbound score of the form:

$$\hat{L}(l_{dn}, l_{dni}, l_{dt}, l_{dti}) \propto f_{dn}(l_{dn}) \cdot f_{dni}(l_{dni}) \cdot f_{dt}(l_{dt}) \cdot f_{dti}(l_{dti}) \cdot \hat{u}\hat{v}\hat{w} \quad (2)$$

This upper bound is now factorizable, and can be used in our A^* search. We can sort all n candidate text groups in an image in decreasing order of $f_{dn}()$, $f_{dni}()$, $f_{dt}()$, $f_{dti}()$, respectively. The resulting four lists contain candidates groups, ordered from best to worst according to our upper bound, for the *DN*, *DNI*, *DT*, *DTI* labels respectively. The top candidates from each list are combined to form our most-promising label-hypothesis.

For the search algorithm we construct a priority queue, designed such that the head of the queue is always the most-promising label-hypothesis yet to be evaluated. A label-hypothesis is evaluated by computing its true score according to formula 1, comparing it to the running best hypothesis (according to true score). The search is over when the true score of the running best hypothesis exceeds the upper bound score of the next most promising candidate waiting at the head of the queue.

Of course, the queue does not have to be populated with all n^4 label candidates at start. It can be grown dynamically as the search proceeds. Whenever we pop the head of the queue, we also insert four successors of the popped label-hypothesis. A successor of a label-hypothesis can be obtained by picking any element of the hypothesis, say the index of the *DN* label – l_{dn}, and replacing it with the next most promising *DN* index, easily obtained by moving down the $f_{dn}()$ sorted list by a notch.

In practice, the queue will seldom process all of n^4 possible label hypothesis, because with a good upper bound, the top label-hypothesis will be found among a few most promising candidates. In our experiments we noted that only of the order of a hundred label-hypothesis candidates were being explored among a potential n^4 candidates, where n was typically between 50 and 200. If the queue for an image grows much larger, it indicates extremely weak discriminating evidence from the unary functions. In such cases our algorithm has a high probability of being in error, even if we let it run till the end, and it may be useful to simply *reject* this input as undecipherable.

This is work in progress. In initial experiments on approximately 1000 test images, the drawing number field was correctly identified about 80% of the time. Observed errors derive from the following major causes: failures of the grouping algorithm to generate correct atomic units of text for labeling; OCR errors; errors in the unary functions to identify the textual properties of target labels.

7 Concluding remarks

Metadata tagging of document content is complementary to brute-force search over unstructured text. With the rise of XML standards for representing metadata in electronic databases, this function is gaining increased societal as well as commercial interest. Discovery of atomic content items, and the tagging of their semantic roles, is a key aspect of this problem.

We have outlined a formal approach to Semantic Role Labeling which decomposes the problem into construction of atomic units using perceptual organization techniques, and subsequent labeling of these units using unary and multiway evidence. We have demonstrated an A* approach to managing inherently exponential search, in cases where the atomic units are fixed. A greater research challenge arises when the atomic units cannot be computed with confidence, but instead multiple candidate parsings must be entertained.

This Semantic Role Labeling approach to document content tagging is amenable to other document recognition data sets and we are interested in expanding the set of domains to which it may be applied.

References

[MV98] E. Miller and P. Viola. Ambiguity and constraint in mathematical expression recognition. In *Proceedings of AAAI-98*, pages 784–791, 1998.

[NKSV93] G. Nagy, M. Krishnamoorthy, S. Seth, and M. Viswanathan. Syntactic segmentation and labeling of digitized pages from technical journals. *IEEE Trans. PAMI*, 15(7):737–747, 1993.

[Sar02] P. Sarkar. An iterative algorithm for optimal style conscious field classification. In *Proceedings of the 16th ICPR*, 2002.

[SM04] E. Saund and J. Mahoney. Perceptual support of diagram creation and editing. In *International Conference on the Theory and Applications of Diagrams*, 2004.

[SN05] P. Sarkar and G. Nagy. Style consistent classification of isogenous patterns. *IEEE Trans. PAMI*, 27(1):88–98, January 2005.

[VN05] S. Veeramachaneni and G. Nagy. Style context with second order statistics. *IEEE Trans. PAMI*, 27(1):14–22, January 2005.

[Wer38] Max Wertheimer. *Laws of Organization in Perceptual Forms (Translation from 1923 German article by W. Ellis)*, pages 71–88. Routledge & Kegan Paul, London, 1938.