Historical Recall and Precision: Summarizing Generated Hypotheses

Richard Zanibbi
Centre for Pattern Recognition and Machine Intelligence
Concordia University, Montreal, Canada
zanibbi@cenparmi.concordia.ca

Dorothea Blostein and James R. Cordy
School of Computing
Queen's University, Kingston, Canada
{*blostein,cordy*}*@cs.queensu.ca*

Abstract

Document recognition involves many kinds of hypotheses: segmentation hypotheses, classification hypotheses, spatial relationship hypotheses, and so on. Many recognition strategies generate valid hypotheses which are eventually rejected, but current evaluation methods consider only accepted hypotheses. As a result, we have no way to measure errors associated with rejecting valid hypotheses. We propose describing hypothesis generation in more detail, by collecting the complete set of generated hypotheses and computing the recall and precision of this set: we call these the 'historical recall' and 'historical precision.' Using table cell detection examples, we demonstrate how historical recall and precision along with the complete set of generated hypotheses assist in the evaluation, debugging, and design of recognition strategies.

1. Introduction

It is important that recognition strategies used in document recognition research be transparent. In particular, a researcher needs to know when hypotheses are created, and how they are modified. This is crucial both for comparing prescriptive recognition theories in experiments, and for detecting errors in strategy implementations (debugging).

We propose a simple approach to increasing the transparency of recognition strategies: record the complete set of hypotheses generated by a recognition strategy, and the history of any rejections and reinstatements of hypotheses. We refer to this record as the *hypothesis history* [10], which we describe in Section 2.

Hypothesis histories provide new information for analysis. In particular, rejected hypotheses which are usually ignored in evaluation are now recorded, permitting new metrics. In this paper we suggest two such metrics for hypotheses generated by a recognition strategy, which we call *historical recall* and *historical precision*. We define these in Section 3, and then discuss various uses of these metrics for evaluation, debugging, and recognition strategy design in Section 4.

We came upon historical recall and precision while considering table recognition research [3, 7, 11] from a decision-making perspective. Reflecting this, we use simple table structure recognition examples for illustration, where cells are detected within a segmented table region.

2. Hypothesis Histories

Baird and Ittner [1, 5] and other researchers including Klein and Fankhauser [6] and Dosch, Rendek, et al. [2, 9] have designed data structures and document recognition frameworks that make it relatively easy to recover intermediate recognition states. Among other benefits, this allows intermediate states to be easily visualized and analyzed. This simplifies debugging a recognition strategy implementation, for example.

Pushing this idea of transparent recognition further, we propose that all unique hypotheses generated should be made available, and that the relations between hypotheses within intermediate states should be collected along with the states themselves. To achieve this, we suggest that the history of each hypothesis needs to be recorded.

A *hypothesis history* [10] describes when hypotheses are first proposed (generated), and the subsequent times at which hypotheses are rejected or reinstated. Reinstatement refers to when a rejected hypothesis is itself rejected, resulting in the hypothesis being accepted again. A hypothesis history also records confidence values associated with hypothesis creation, rejection, or reinstatement (e.g. probabilities or fuzzy values). For brevity, in the remainder of this paper we will treat hypotheses as being either true or false.

We illustrate hypothesis histories in Figure 1, using a simple cell detection example. Shown are the sets of accepted and rejected hypotheses as a cell detection strategy

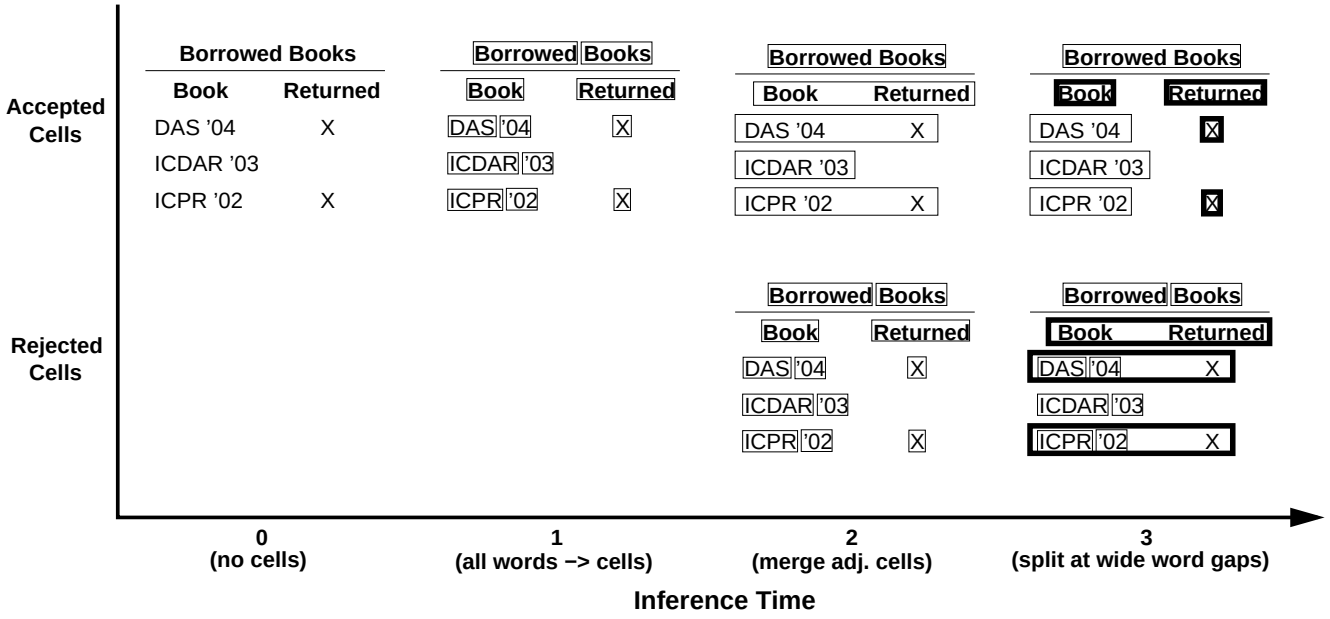

Figure 1. Accepted and rejected hypotheses of a table cell detection strategy over time

progresses. The points at which hypotheses are proposed or have their truth value altered by a strategy are referred to as *inference times*, and are indicated on the horizontal axis in Figure 1.

At inference time 0, no cell hypotheses have been proposed; we show the original table image here to assist the reader. At inference time 1, all words are proposed as new cell hypotheses. At time 2, all horizontally adjacent cells are merged to create an entirely new set of cell hypotheses. This results in all cell hypotheses from time 1 being rejected. Finally, at time 3 a correction is made by splitting cells proposed at time 2 at wide word gaps (this might be defined by a threshold, for example). Three of the cell hypotheses from time 2 are rejected, and four of the initial cell hypotheses from time 1 are reinstated. These changes in truth-value are indicated using thicker box outlines at time 3.

Note that the hypothesis reinstatements at inference time 3 in Figure 1 are implicit. They occur not because their rejection was reconsidered directly, but because splitting the accepted cells from inference time 2 produced previously rejected hypotheses that were originally proposed at time 1.

Conventionally the analysis of recognition results is carried out using only hypotheses accepted at the final inference time. In Figure 1, this would be the set of accepted cells at time 3. This 'black box' view of the recognition strategy's progress disposes of significant information. For example, the set of cells that have been correctly proposed but rejected are not described. If intermediate states or rejected hypotheses are considered when analyzing recognition results, it is usually done informally within the context of an error analysis.

In contrast, we can obtain a 'clear' or 'white' box view of the strategy's progress by recording a hypothesis history. We can then incorporate the set of all generated hypotheses into evaluation and error analysis using well-defined metrics and automated tools. In the next section we will define two metrics for summarizing hypotheses generated by a recognition strategy.

Elsewhere we have described a simple graph data structure for capturing hypothesis histories, and a strategy specification language which supports the automatic recording of hypothesis histories during the execution of a strategy [10]. Systems which already record intermediate states, such as those mentioned at the beginning of this section, can probably be modified to record hypothesis histories with relatively little effort.

3. Historical Recall and Precision

If we record the history of hypothesis creation, rejection, and reinstatement produced by a recognition strategy, we are able to observe some new metrics. We will now define and describe two such metrics for the set of hypotheses generated by a strategy, which we have named *historical recall* and *historical precision.*

As could be seen in Figure 1, at a given inference time the set of generated hypotheses (e.g. cell locations) is defined by the union of hypotheses that are currently accepted (A) and rejected (R). The validity of individual hypotheses within A and R is determined by the set of recognition tar-

gets (T), often referred to in the document recognition literature as *ground truth*.

As shown in Figure 2, correct hypotheses (C) are defined by the intersection of accepted hypotheses (A) and the set of recognition targets (T). Similarly, rejection errors (correct hypotheses that have been rejected, F) are defined by the intersection of rejected hypotheses (R) and recognition targets (T).

Figure 2 also presents a number of metrics. *Conventional* recall and precision describe the ratio of correct hypotheses (C) to recognition targets (T) and accepted hypotheses (A) respectively. Consider inference time 2 in Figure 1. Assuming that the eight cells accepted at inference time 3 comprise the recognition targets (T), then two of the five accepted cell hypotheses are correct at time 2 ($|A| = 5$, $|C| = 2$). Recall is then $2/8$ (25%), and precision is $2/5$ (40%) at inference time 2.

Historical recall and precision describe the recall and precision of the complete set of generated hypotheses. The set of generated hypotheses is defined by the union of accepted and rejected hypotheses ($A \cup R$), while the the set of generated hypotheses matching recognition targets are defined by the union of correct and falsely rejected hypotheses ($C \cup F$). If no hypotheses are rejected (i.e. R is empty), then the 'conventional' and historical versions of recall and precision are the same. The key difference here is that the historical metrics take rejected hypotheses into account, while the conventional ones do not.

Again using inference time 2 in Figure 1 as an example, there are twelve rejected cell hypotheses ($|R| = 12$), but with four rejected incorrectly ($|F| = 4$; these are the cells with thick boxes accepted at inference time 3). Incorporating $|A|$, $|C|$, and $|T|$ as computed above, the historical recall is then $(|C| + |F| = 6)/8 = 75\%$, and the historical precision is $(|C| + |F| = 6)/(|A| + |R| = 17) = 35.3\%$ at inference time 2.

Conventional and historical recall may be directly compared, as they both describe coverage of the set of recognition targets. Note that historical recall will always be greater than or equal to recall. The difference of historical and conventional recall is the proportion of recognition targets that have been falsely rejected (shown as *Rejected Targets* in Figure 2). For inference time 2 from Figure 1, the rejected target ratio is $(|F| = 4)/(|T| = 8) = 50\%$; exactly half the target cells have been proposed and rejected.

It is harder to directly compare conventional and historical precision. This is because they do not describe proportions of the same set: conventional precision describes the proportion of accepted hypotheses (A) that are correct, while historical precision describes the proportion of accepted and rejected hypotheses ($A \cup R$) that are correct.

The additional information provided by historical precision may be understood as the accuracy of hypothesis generation. Choosing between two strategies, a designer might consider the cell detection strategy with the higher historical precision to be more elegant, because it considers fewer incorrect possibilities.

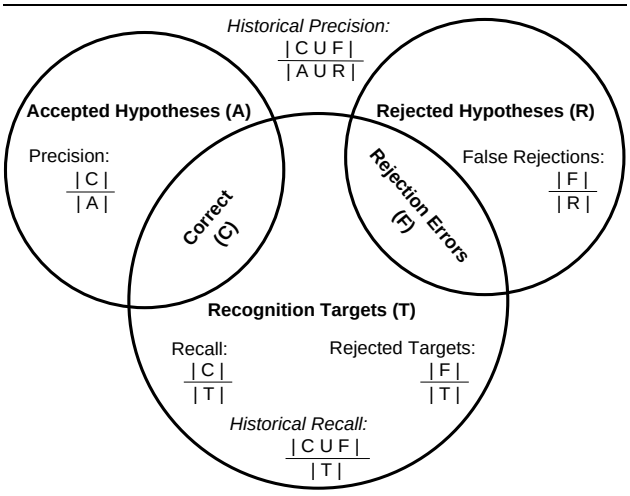

Figure 2. Venn diagram illustrating recall, precision, historical recall, and historical precision

Recall and precision metrics may be modified to allow approximate matching between accepted hypotheses (A) and recognition targets (T). This modifies the magnitude of C, and thus recall and precision. In the table recognition literature, approximate matching has been achieved through confidences (e.g. based on words in a cell region) and edit distances [11]. Approximate hypothesis matches are represented as values in the interval [0,1], to describe the closeness of a match. One could also apply this approach to define approximate matching between rejection errors (F) and recognition targets (T), modifying historical recall and precision values in the process.

Alternatively, hypotheses with associated probabilities or confidences may have their truth values binarized (as accepted/rejected) by rejecting hypotheses with a confidence below a threshold value. This produces the accepted (A) and rejected (R) hypothesis sets needed to compute the metrics shown in Figure 2.

4. Analysis Using Hypothesis Histories, and Historical Recall and Precision

Figure 3 presents results for accepted and rejected cell hypotheses of the Handley table recognition strategy [4] over time, for the challenging table shown. The table is taken from the University of Washington Database [8], page a038. We provided the Handley strategy with lines and bounding boxes for words located within table cells as in-

Table 49.—Average values for bulk density, grain density, and total pore space of gray dacite from the lateral-blast deposits and of pumice lapilli from pyroclastic-flow deposits of Mount St. Helens

Type of deposit	Bulk density Mean (g/cm^3)	Bulk density No.[1]	Grain density Mean (g/cm^3)	Grain density No.[1]	Total pore space (percent)
Lateral blast, May 18--------	[2]1.66	262	2.52	3	36
Pyroclastic flow,					
May 18-------	.74	8	2.55	3	[3]71
May 25-------	.95	2	(3)	0	[3]63
June 12------	1.08	10	2.53	3	57
July 22------	.88	11	2.55	1	65
August 7-----	1.02	12	2.61	3	61
October 16-18	1.12	12	2.65	5	58

[1] Number of determinations.
[2] Data from Hoblitt and others (this volume).
[3] Grain density (Dg) not determined; total pore space calculated using Dg=2.60.

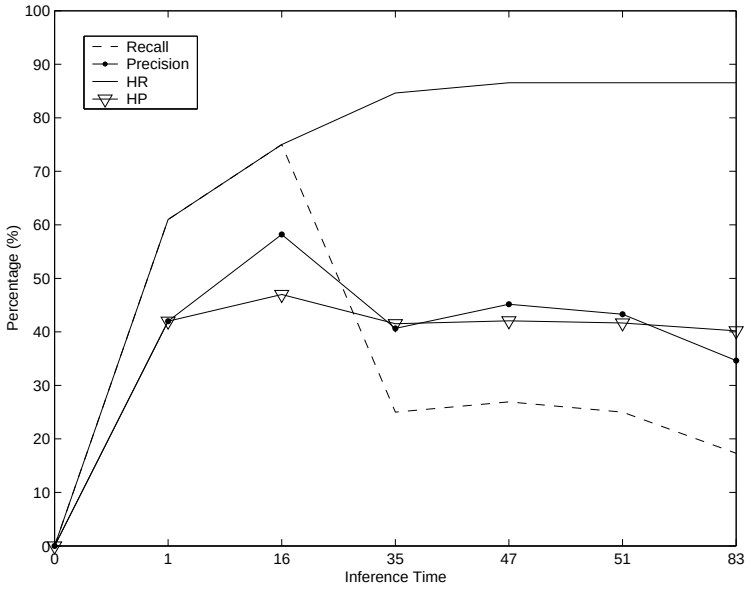

Figure 3. Cell detection results over time (right) for a text table (left). Inference times shown are those affecting cell hypotheses. *HR* is historical recall, and *HP* historical precision

put. Correct cells were defined as word sets by one of the authors (this defined the recognition targets, T in Figure 2).

Shown on the right side of Figure 3 are the recall, precision, historical recall, and historical precision of accepted and/or rejected hypotheses at each inference time when cell hypotheses were modified. The inference times correspond to points where a decision function returns a result, reflecting the specification language that was used to reimplement the Handley strategy [10].

We can see in Figure 3 that recall and historical recall are identical until inference time 35, when a large number of correct cell hypotheses are rejected due to a decision to merge cells, decreasing recall sharply. Roughly half of the correct hypotheses at inference time 16 are rejected at time 35. Historical recall increases at time 47 and then levels off, while recall increases slightly at time 47 and then drops to it lowest point at time 83. In the end, nearly 90% of the cell targets are hypothesized (the historical recall), but less than 20% are accepted in the final result.

Throughout the inference times shown in Figure 3, precision and historical precision are roughly identical. Conventional and historical precision are not directly comparable, as discussed in the previous section. We can observe that never more than 50% of the generated cell hypotheses are correct, as indicated by historical precision. This partly reflects that the Handley strategy initially proposes that all words are cells (just as in Figure 1, time 1). Conventional precision is highest at times 16-34 (over 55%), and lowest in the final result (less than 40%). Historical precision allows us to observe that in addition to many spurious cells being accepted in the output (low precision), many spurious cells have been generated.

The hypothesis history that allows us to compute the historical recall and precision values shown in Figure 3 also permits us to determine exactly which hypotheses are in the accepted and rejected sets at each inference time, and how hypotheses move between the sets through time.

We can also determine how *decisions* (e.g. individual classification or segmentation results) cause changes at each inference time, simply by adding this extra bit of information to our hypothesis history. For example, we were able to determine that the slight increase in all four metrics at time 47 was due to the header cell ‘Total pore space (percent)’ being created by merging two cells containing ‘Total’ and ‘pore space (percent)’ separately.

The additional information provided by hypothesis histories allows a person designing or debugging a strategy implementation to quickly locate weak decisions and examine their effects. This can be done informally examining plots such as the one shown in Figure 3, or formally using descriptive statistics and simple tools that search through hypothesis histories.

For evaluation, in addition to making the progress of individual strategies transparent, hypothesis histories allow strategies to be compared by their generated hypothesis sets ($A \cup R$ in Figure 2). In an evaluation of cell detection strategies, we could determine which cell hypotheses were proposed by all strategies, and which were considered by only

one strategy. Similarly, we could determine which cell hypotheses were rejected by all strategies, some of which may be recognition targets.

In summary, historical recall and precision provide a high-level summary of decision making in recognition strategies, while hypothesis histories provide a detailed low-level view. Together they allow strategies to be analyzed and compared with greater precision than in the current common practice, where hypotheses rejected in the output are ignored.

5. Conclusion

We have proposed that the decision process of document recognition strategies be made more transparent by recording the creation, rejection, and reinstatement of hypotheses. These *hypothesis histories* allow us to take rejected hypotheses into account. Further, they allow us to determine the exact effects of individual decisions on hypotheses, as discussed in the previous section. This is quite useful for strategy design, debugging, and evaluation.

Given a hypothesis history, we can compute *historical recall* and *historical precision.* Historical recall is the proportion of recognition targets proposed as a hypothesis (e.g. for table cell detection, correct cell regions). Historical precision is the proportion of generated hypotheses that match recognition targets. In Section 3 we discussed how historical recall and precision complement conventional recall and precision in analysis.

In addition to historical recall and precision, additional metrics based on hypothesis histories are worth exploring in the future. For example, one might define the 'fickleness' of a strategy as some function of the number and/or frequency of hypotheses moving between the sets of accepted and rejected hypotheses. These new metrics would provide additional means for summarizing and understanding the decision making behaviour of recognition strategies.

Acknowledgments

We wish to thank Dr. C.Y. Suen and CENPARMI for providing the resources to write this paper. This research was funded by the Natural Sciences and Engineering Research Council of Canada.

References

[1] H. Baird and D. Ittner. Data structures for page readers. In A. Spitz and A. Dengel, editors, *Document Analysis Systems*, pages 3–15. World Scientific, Singapore, 1995.

[2] P. Dosch, K. Tombre, C. Ah-Soon, and G. Masini. A complete system for analysis of architectural drawings. *Int'l J. Document Analysis and Recognition*, 3(2):102–116, Dec. 2000.

[3] J. Handley. Document recognition. In E. Dougherty, editor, *Electronic Imaging Technology*, pages 289–316. IS&T/SPIE Optical Engineering Press, Bellingham (USA), 1999.

[4] J. Handley. Table analysis for multi-line cell identification. In *Proc. Document Recognition and Retrieval VIII (IS&T/SPIE Electronic Imaging)*, volume 4307, pages 34–43, San Jose (USA), 2001.

[5] D. Ittner and H. Baird. Programmable contextual analysis. In A. Spitz and A. Dengel, editors, *Document Analysis Systems*, pages 76–92. World Scientific, Singapore, 1995.

[6] B. Klein and P. Fankhauser. Error tolerant document structure analysis. *Int'l J. Digital Libraries*, 1(4):344–357, Dec. 1997.

[7] D. Lopresti and G. Nagy. Automated table processing: An (opinionated) survey. In *Proc. Third Int'l Workshop Graphics Recognition*, pages 109–134, Jaipur (India), 1999.

[8] I. Phillips, S. Chen, and R. Haralick. CD-ROM document database standard. In *Proc. Second Int'l Conf. Document Analysis and Recognition*, pages 478–483, Tsukuba Science City (Japan), 1993.

[9] J. Rendek, G. Masini, P. Dosch, and K. Tombre. The search for genericity in graphics recognition applications: Design issues of the Qgar software system. In *Proc. Sixth Int'l Workshop Document Analysis Systems*, pages 366–377, Florence (Italy), Sept. 2004.

[10] R. Zanibbi. *A Language for Specifying and Comparing Table Recognition Strategies*. PhD thesis, Queen's University, Kingston (Canada), Dec. 2004.

[11] R. Zanibbi, D. Blostein, and J. Cordy. A survey of table recognition: Models, observations, transformations, and inferences. *Int'l J. Document Analysis and Recognition*, 7(1):1–16, Sept. 2004.