



## Best Practices: Improving ICR Accuracy with Better Form Design

---

By Rick Scanlan

Throughout time people have been looking for ways to gather information. The invention of the printing press allowed large scale production of documents and printing of the earliest forms. Until the 1980's information collected from forms was tabulated by hand or manually entered into a computer. Hand print recognition technology has progressed significantly since that time, but the accuracy and productivity of forms processing is highly dependent on form design.

There are many factors to consider when designing a form to collect hand printed responses. The form needs to be easily understood by your target audience. The form needs to make very clear where the user is to write their responses, and the response area should be constrained.

Remember that the person filling out the form is usually out of your control. No matter how well you design your form there will always be responses that can't be read automatically. You can encourage the form fillers to write neatly, and keep their responses within the spaces allotted, but there will always be people who don't read instructions (or don't care) and assume that the form will be read by a human, not by a computer. They do things like writing a character by mistake then drawing a big "X" over it to "delete" it. Some people just have really poor handwriting, or always write in cursive.

The industry average for Intelligent Character Recognition (ICR) is about 70%. That means that three out of every ten characters are read incorrectly or aren't recognized with a high enough confidence to be considered accurate. One should never expect 100% accuracy in any forms processing project. A successful ICR application should exceed 70% accuracy. 85% or higher is considered good. (That's still 15 bad characters out of every 100.) With a little planning and some basic form design elements one can usually exceed the 70% threshold.

A great example of successful ICR forms applications is the tax forms used by the State of Oklahoma. Oklahoma has been successfully using ICR forms since the early 1990's. Their success is due to several factors, including excellent form



design elements plus the importance of the document to the users completing the forms.

## **General Form Design Considerations**

There are a few easy things that can be done to improve ICR accuracy. First, tell the user that the form will be processed by a computer. Stress the importance of writing plainly, carefully and clearly. Ask them to use block letters. Put the instructions in bold at the top of the form, or just above the first field. Show character examples such as how an “A” or a “2” should be formed. Simple user instructions without changing any other aspect of the form can provide a significant improvement in recognition rates.

## **Field Design**

Properly laying out the areas for printed responses can make the most significant impact in accurate hand print content recognition. A common mistake in field design is to provide a free form area for a response. This design is often a simple blank line where people should write. Without any character restraints people will write in cursive, will run their characters together, will write on top of the line, and will write multiple lines in a single line response area. All of these factors will have a serious impact on recognition accuracy.

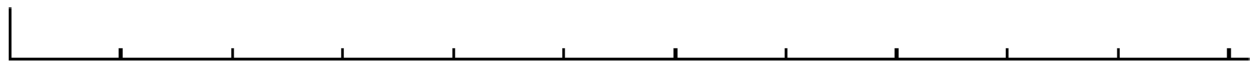
A form needs to have a defined response area for each character, encouraging character separation. Some approaches for character separation work better than others, and are described below.

### Comb Lines

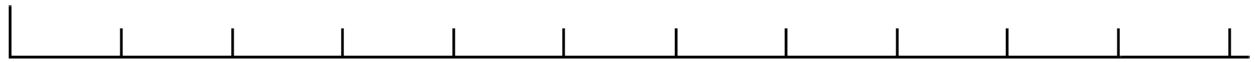
Comb lines are horizontal lines with small vertical separators called tick marks. This is traditionally the most common type of hand print form design, often used in manual data entry applications. However, it is not as well suited for automated ICR processing as other approaches. While the tick marks may encourage people to separate their characters they rarely ever write the characters within each space. The spacing between the vertical lines on many forms is frequently too close together, making it almost impossible for the average person to stay between the lines. The height of the tick lines also plays an important role in encouraging character separation.

If you use comb lines, provide plenty of room between each vertical tick marks. Make the tick marks tall enough to encourage people to write between them. A

vertical height at least half the height of the expected character is usually sufficient.



Example of a Poor Comb Line



Example of a Good Comb Line

### Character Boxes

Character boxes are usually the best method to encourage character separation. A good character box design will allow users to write their characters completely within each box. Unfortunately many forms contain boxes that are too small and too close together. People often can't write small enough to keep an entire character within a box. Pencil lead creates strokes that are usually much wider than with pens, making it even harder to constrain the character. The following are some general guidelines for designing character boxes.

Each box should be square in shape. Rectangular boxes with the height taller than the width can make the user feel like they need to squeeze their characters into the space. This often results in characters written in a compressed vertical form, reducing accuracy. A square shape encourages wider, more normally formed characters.



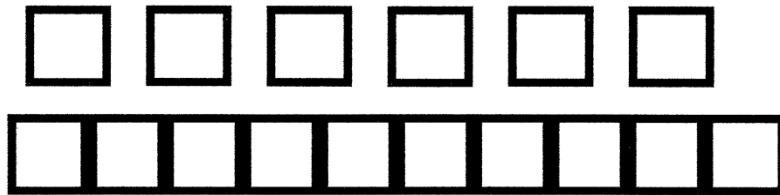
Narrow Boxes



Square Boxes

Single character response locations such as for Male ("M") or Female ("F") should be provided in a single box separated from other responses.

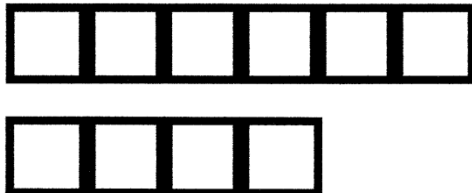
Multiple character response locations such as a Name field may contain separated boxes if space permits, or may be joined together. If joined, a thick separator between the response locations, at least one-fourth the width of a response area, should be used to discourage characters entering other boxes.



Individual fields should be separated by enough space to easily identify where one field stops and the next starts. Spacing of at least 1.5 box widths is recommended to prevent users from interpreting the space as a valid character location.



Rows of fields stacked vertically should be separated by at least one half the height of an individual box.



Boxes can be printed with either solid black lines or dropout colors, depending on the scanning and forms processing technology used. Some forms processing technology such as the [FormSuite](#) Software Development Kit (SDK) from Accusoft Pegasus best performs form identification and alignment when all boxes and form contents are retained. Software-based form dropout is used to remove the boxes from the image after scanning. The original image with boxes intact may then be archived for future reference.

Some forms processing systems require dropout colors to be used when printing forms. For example, a form is printed in red ink, and a red bulb in the scanner eliminates the red content when the image is captured. Other forms processing technology, such as the FormSuite SDK provided by Accusoft Pegasus, does not require any special printing, paper or inks. This type of technology can provide much greater printing flexibility and reduced printing costs. The use of general purpose scanning technology without requiring special bulbs may also reduce capital costs.

### **Paper Thickness and Bleed-Through**

The quality of paper can impact recognition accuracy in dual-sided forms. Form paper should be thick enough to prevent the back side content from bleeding through when scanned. Fields on form fronts and backs may also be offset to ensure that any bleed-through content from one side will not interfere with field recognition on the other.



Example of Bleed-Through

### **Processing a Hand Print Form**

Recognition accuracy can often be increased through image enhancement and other pre-processing activities. Many scanners today include image enhancement technologies that will create a good representation of the original image. That image works well for viewing or archival. However, that image may not be the best to use for content recognition. Lines or boxes in the image may interfere with field recognition. Dot shaded fields may prevent easy recognition of filled content. Filled forms might be received via fax at a low resolution, where built-in image enhancement is not available. The use of post-scan image enhancement processes can significantly improve forms processing and content recognition.

A temporary copy of the image can be created solely for use in content recognition. Enhancements are performed that directly impact recognition. If poor recognition results are received, additional enhancements may be performed,



looping through a series of “enhance – attempt to recognize – enhance – attempt to recognize” processes until the field is read with high confidence or a decision is made to route the image for manual data entry. Once recognition is complete the temporary image is deleted and the original image is archived.

Certain enhancement processes are specifically designed to improve character recognition, especially when you don’t have control over the form design. For example, forms that contain shaded fields in response areas can be very difficult to recognize. Dot shading removal with character smoothing can significantly improve recognition of those fields.

Software-based form dropout, removal of background form content, can allow recognition of content that has been written over master form elements. For example, users completing forms with comb lines often write on top of the lines, resulting in very difficult recognition. An automated comb line removal process, such as that provided by the ScanFix technology from Accusoft Pegasus, will remove the comb lines and reconstruct the intersecting characters, allowing for accurate recognition.

~~718-412-711~~

Before Comb Removal

78-4271

After Comb Removal, with Character Repair

### Focused Recognition and Data Validation

Some fields are designed to allow only certain characters to be entered. For example, a date field may allow only digits, or only digits, dashes and slashes. A “Male/Female” field may only allow the characters M and F. Ensure that your form contains instructions or examples for each field to ensure the user knows what characters are allowed. ICR technology such as Accusoft Pegasus’ SmartZone ICR / OCR component allows definition of allowable characters, increasing accuracy by focusing the recognition engine towards specific characters.

Remember that the industry average for hand print recognition is only 70%. Data validation and correction is critical to a successful hand print forms recognition system. Use recognition confidence values to locate suspect characters. Use two or more ICR engines in a voting process, comparing the results from each engine to determine the highest confidence results. Recognized data should be



compared against database tables, dictionaries, lookup tables or other data validation tools.

A “key from image” process is typically required to validate low confidence data. You should plan for development of a process to display suspect characters or fields to a human for manual data entry. Human interaction is the most expensive part of any data capture process, so any efforts you can take, such as strong form design or additional image enhancement processes, will easily pay for themselves when compared to the cost of manual data entry.

### **Test Your Form**

You should develop a prototype of your form then test it on a sampling of actual users. Present the form for completion by people who have not previously seen any version of form, and ask them to complete it. Statistical sampling and analysis is helpful when testing forms that will be used on a large scale. Forms for smaller audiences do not require scientific analysis. Just be sure that representative users are used in the test.

You should also test your recognition processes with enough sample data to get a good sense of the results. Identify weaknesses in the form or recognition, make changes, then retest to confirm improved results.

### **A Note About OMR**

Optical Mark Recognition (“OMR”), sometimes known as “mark sense,” is the analysis of form locations to determine if a mark is present. Examples of OMR zones include check boxes on a form to designate male or female, multiple choice responses on a high stakes educational exam, or diagnosis results on a medical form. OMR response areas may be a single box, multiple response zones such as a “check all that apply” field, or a true/false designation. Many forms contain some type of OMR field. Designing an OMR field is simpler than for character recognition but still requires careful consideration. Whether an oval bubble, square box or open brackets are used, be sure the area is large enough for the user to easily mark within the designated area. Common OMR field design errors includes making the box too small for people to easily mark within the zone, or printing the boxes too close together, resulting in more than one box containing the mark. Some users will circle an OMR response area instead of filling in the box. Similar to character responses, providing clear instructions and example marks can significantly improve recognition results.



Even great instructions will not prevent some people marking a zone in error then drawing a big “X” over in an attempt to “delete” the mark. Business rules must be developed to handle multiple mark situations and manual key-from-image operations are usually required to determine user intent.

## **In Conclusion**

Many factors influence the accuracy and success of a hand print forms processing system. Extra time and consideration spent in forms design will pay strong dividends in recognition accuracy and reduced costs for many data entry. Carefully consider your target audience and design a form that will be easily understood and completed, and can be easily recognized and processed.

You can find Accusoft Pegasus product downloads and features at [www.accusoft.com](http://www.accusoft.com). Please contact us at [sales@accusoft.com](mailto:sales@accusoft.com) or [support@accusoft.com](mailto:support@accusoft.com) for more information.

## **About the Author**

*Rick Scanlan, Director of Sales Engineering*

Rick Scanlan joined Accusoft Pegasus with the acquisition of TMSSequoia in December 2004 after 15 years of service. During his tenure, Rick has served in a variety of technical, business development and corporate management roles including technical consultant for TMS' early CD-ROM publishing products, business development for imaging tools, sales management, and corporate management. Over the years, Rick has developed extensive expertise in a wide variety of imaging technologies including Internet-based imaging, image enhancement and forms processing. As General Manager of TMSSequoia's Internet Innovations Division, he was responsible for directing engineering, sales, and customer support. Rick was instrumental in bringing the Prizm® Viewer to market. Rick currently utilizes his years of imaging and consulting experience to assist the sales team in analyzing customer technical requirements and assisting with sales activities. Rick also helps define Accusoft Pegasus product strategy and future development. A native of Oklahoma, Rick earned Bachelor of Science degrees from Oklahoma State University in Business Management, Economics, and Management Science and Computer Systems.