Top Fifteen Basic Tips

- Pages should be clean with ample white space or separator lines between areas
- Page layout should exist without heavy decorative graphics or transparency
- Avoid ambiguous layouts or ordering as there is no semantic reasoning component
- Subtractive layout (white rects or text on top of colored rects) should be avoided
- Be consistent in the representation of various field types (don’t mix and match styles)
- Do not have overlapping objects where you expect a field to be (esp. underlines)
- Don’t be tricky or clever in creation of field areas (no wide strokes instead of fills)
- Try to ensure clean intersections at corners (not too much overshoot)
- Ideally use circular objects for radio buttons and square objects for check boxes
- Always use lined tables with top header cells having full-justified text inside
- Use regular comb fields (do not vary the dimension of individual cells or spacing)
- Try to keep text within a reasonable size range (10 point to 24 point) for best results
- Avoid drop shadows and other decorative flourishes on boxes and text
- Make sure fonts used have proper encodings (especially for Ascii art characters)
- Do not share labels across multiple fields when possible (use separate labels)

The Basic Process

- If the page is image-only, it is fed to OCR and objects grabbed directly from there
  - Lines of text including special characters (boxes, circles, horizontal lines)
  - Horizontal and vertical lines, stroked rectangles
- If normal PDF, look at graphic objects, text, and current annotations
  - Lines of text including special characters (boxes, circles, horizontal lines)
  - Horizontal and vertical lines, rectangles (hollow and filled)
  - Diagonal lines only if they form a diamond-shaped appearance
  - Curved lines if and only if they form circular appearance
  - Annotations are noted so as to not create duplicates and for ordering
- A multi-step process is used to find form fields of the type:
  - Check boxes
  - Radio buttons
  - Digital signature fields
  - Underline and box text fill-in
  - Open and close comb fields
  - Tables of text fill-in fields
- Labels are assigned to each type of object in a multi-step process
  - Table labels based on information inside the cell or at top and left headers
  - Find the highest number of matches in a direction for a given form field type
  - Keep repeating above step until all individual field components have been found
  - Lastly look for group labels (i.e. radio buttons - required)
- Create the new form fields in the PDF
  - Ensure that the tab order is correct based on reading order of form
  - If tagged document, add new annotations to the structure tree
Specifications for Optimal Recognition

• Stroked square-, circular-, and diamond-shaped check boxes
  o Should be symmetric about x- and y-axis
  o Should be monotonically increasing or decreasing in x and y
  o Can be special characters (e.g. ZapfDingbats, Wingdings, and WebDings)
  o Unicode appearances of rectangle, circles, or diamonds
  o Labels usually to the right

• Stroked circular-shaped or numbered radio buttons
  o Must have a label for the radio button group (or will default to check box group)
  o Grouping must be a single row or column of circular or numbered buttons
  o Individual button labels to the right; group label to the left or above

• Digital signature fields
  o Starts off as a normal text fill-in
  o Promoted only if “Signature” seen – not localized

• Underlines for text fill-in so long as they follow a regular pattern
  o Absolutely no text may intersect the bounds of the underline in x-direction
  o Upper bounds based on intersection with other objects up to ~36-point maximum
  o Underlines can be solid, dashed, dotted, or a dash-dot pattern
  o Can be graphic lines or the underscore, dash, and/or dot characters
  o Do not have dashes or slashes on top of the underlines (break them up)
  o Can have multiple fields for single underline if well-separated labels underneath
  o Labels usually to the left or below

• Singleton hollow boxes
  o Solid lined borders must not intersect any other graphic objects
  o Label usually comes from inside, or is taken from left or above
  o Multi-line fields are recognized if box is taller than ~36-point

• Dual horizontal boxes
  o Must have text in the left cell and empty in the right

• Closed comb fields (i.e. top line is drawn)
  o Cells can be touching or merely adjacent
  o Cells should not be excessively wide compared to height
  o Label taken from above, below, or to the left

• Open comb fields (i.e. no line on top)
  o Cells must have a U shape if not touching not L
  o Tick marks in between characters must be of the same height
  o Label taken from above, below, or to the left

• Table cells if text not present in the cell
  o Must have a header cell containing text in the top row for each column
  o All cell borders must be drawn lines (except extrema)
  o Adjacent cells must share the same border line
  o Can have text inside a data cell if placed in the top or left portion
  o Secondary direction (left or top) should be left, right or center justified
  o Headers should have no unused space larger than dominant text size of table
  o Labels are based on the names of the row and column headers
  o If row or column header label is not available, will use Column_N or Row_M respectively