Introduction:

The problem involves extracting structured text from images where headings and subheadings are visually distinguished by colour, indentation, and text size. The challenge is to identify and extract the text while simultaneously recognizing and classifying headings and subheadings based on these visual cues. The goal is to convert visual data into a machine-readable, organized format.

Objective:

The aim is to develop a solution that accurately extracts text from an image and classifies the text into headings and subheadings during the extraction process. This will be done by leveraging visual features such as colour, indentation, and text size, allowing for the text to be output in an organized dictionary structure.

Proposed Solution:

The solution will utilize three main methods, all applied while extracting the text, to distinguish between headings and subheadings:

1. **Colour-based Recognition**:
    - **During text extraction**, we will use colour segmentation to differentiate between headings and subheadings. Since headings and subheadings are often coloured differently, we can classify them based on their respective colours by segmenting the image into different colour regions using OpenCV's colour detection methods.
2. **Indentation-based Recognition**:
    - **While extracting text**, the spatial layout of text will be analyzed by detecting the positions and alignments of text blocks. Indented text typically indicates subheadings, and this can be identified based on the relative positions of bounding boxes around the text.
3. **Text Size-based Recognition**:
    - **During extraction**, text size information will be used to classify larger text as headings and smaller text as subheadings. Bounding box dimensions or font size metrics will be used to measure and differentiate the size of the text elements.

Solution Steps:

1. **Image Preprocessing**:
    - Convert the image to a suitable format for text recognition (grayscale, thresholding) while maintaining colour information for colour-based discrimination. This step ensures that the OCR system can accurately read the text while also preserving necessary visual distinctions for classification.
2. **Text Extraction and Classification**:
    - **Colour-based Text Extraction**: Use colour segmentation to group text by colour, helping to classify headings and subheadings during the OCR process.

- **Indentation-based Extraction**: Use bounding box detection to understand the alignment of text and identify indented subheadings relative to their corresponding headings.
- **Font Size-based Extraction**: Measure the size of bounding boxes around the text during extraction to differentiate larger text (headings) from smaller text (subheadings).

3. **Organizing Data**:
    - As the text is extracted and classified based on colour, indentation, and size, it will be directly structured into a dictionary format. Headings will serve as dictionary keys, and the associated subheadings will be stored as the corresponding values. The three methods (colour, indentation, and size) will ensure accurate classification and organization of text.

Technologies and Tools:

- **Tesseract OCR**: For extracting text from the image while simultaneously using the layout information provided by the OCR engine.
- **OpenCV**: For colour segmentation, text block detection (bounding boxes), and identifying indentation and size characteristics.
- **Python**: To implement the solution, using libraries like tesseract for OCR, cv2 for image processing, and custom code to organize the extracted data.

Expected Outcome:

The solution will output an organized dictionary where headings are correctly mapped to their corresponding subheadings, achieved by analysing visual cues such as colour, indentation, and text size during the extraction process. This method will ensure accurate recognition of structured information from complex images.