ACCirO: A System for Analyzing and Digitizing Images of Charts with Circular Objects[*]

Siri Chandana Daggubati and Jaya Sreevalsan-Nair[**][0000−0001−6333−4161]

Graphics-Visualization-Computing Lab (GVCL),
International Institute of Information Technology, Bangalore (IIITB),
26/C, Electronics City, Karnataka 560100, India
{daggubati.sirichandana,jnair}@iiitb.ac.in
http://www.iiitb.ac.in/gvcl

Abstract. Automated interpretation of digital images of charts in documents and the internet helps to improve the accessibility of visual representation of data. One of the approaches for automation involves extraction of graphical objects in the charts, *e.g.,* pie segments, scatter points, etc., along with its semantics encoded in the textual content of the chart. The scatter plots and pie charts are amongst the widely used infographics for data analysis, and commonly have circle objects. Here, we propose a chart interpretation system, ACCirO (Analyzer of Charts with Circular Objects), that exploits the color and geometry of circular objects in scatter plots, its variants, and pie charts to extract the data from its images. ACCirO uses deep learning-based chart-type classification and OCR for text recognition to add semantics, and templatized sentence generation from the extracted data table for chart summarization. We show that image processing and deep learning approaches in ACCirO have improved the accuracy compared to the state-of-the-art.

Keywords: Image processing · Scatter plots · Pie charts · Dot plots · Bubble plots · Circle geometry · Circle Hough Transform (CHT) · Spectral clustering · Chart data extraction · Text recognition

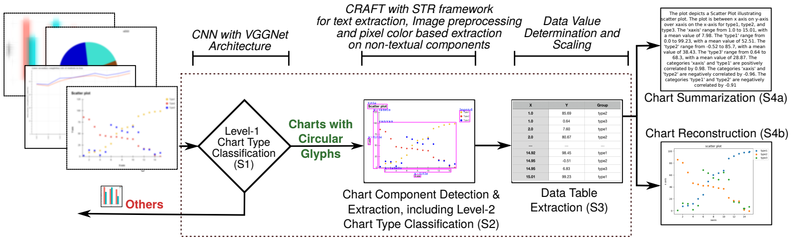
1 Introduction

Given the ubiquity of charts for visualizations, there is recent interest in automating its interpretation. The motivating applications include filtering significant charts from image databases, generating visual question-answering (QA) systems, etc., given the raster format of the charts. Though many image processing techniques have been used, there are still gaps in existing technology for automated chart interpretation owing to the diversity and complexity of chart content and the requirement of human-in-the-loop in most cases. Here, we

[*] This work has been generously supported by the Machine Intelligence and Robotics (MINRO) Grant, Government of Karnataka. This paper has benefitted from discussions with GVCL members and peers at IIITB.

[**] Corresponding author

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_50


CNN with VGGNet Architecture
CRAFT with STR framework for text extraction, Image preprocessing and pixel color based extraction on non-textual components
Data Value Determination and Scaling
Level-1 Chart Type Classification (S1)
Charts with Circular Glyphs
Others
Chart Component Detection & Extraction, including Level-2 Chart Type Classification (S2)
Data Table Extraction (S3)
Chart Summarization (S4a)
Chart Reconstruction (S4b)
scatter plot
type1
type2
type3

The plot depicts a Scatter Plot illustrating scatter plot. The plot is between x axis on y-axis over xaxis on the x-axis for type1, type2, and type3. The 'xaxis' range from 1.0 to 15.01, with a mean value of 7.98. The 'type1' range from 0.0 to 99.23, with a mean value of 52.51. The 'type2' range from -0.52 to 85.7, with a mean value of 38.43. The 'type3' range from 0.64 to 68.3, with a mean value of 28.87. The categories 'xaxis' and 'type1' are positively correlated by 0.98. The categories 'xaxis' and 'type2' are negatively correlated by -0.96. The categories 'type1' and 'type2' are negatively correlated by -0.91

Fig. 1: Our proposed workflow for ACCirO is an automated system for data extraction from images of charts with circular objects and color, and for which geometric and color-based information extraction techniques are used.

consider pie charts, scatter plots, and its variants, *dot plots*, and *bubble plots*, with circular objects. The source images are from documents in portable document format (.pdf), websites, outputs from plotting tools, and curated image databases [7]. The data extracted from images comes from its non-textual and textual content in charts. The non-textual content implies the geometric objects as per chart type, *e.g.,* pie segments and scatter points. The text content is from chart, legend, and axes titles, which are localized using annotation and extracted using image processing and text recognition.

When using the second-order gradient tensor field-based approach for object extraction from charts [4, 5], we observe that only regions with high color gradients (edges and corners) are extracted, *e.g.,* boundaries of bar objects. But in pie charts, the gradients are concentrated in the corners of the largest rectangle enclosed within the pie owing to the high curvature gradient. At the same time, the pie chart has a circle geometry that can be exploited for sector extraction. Thus, we propose using color and geometry information in pie charts for automated annotation and data extraction. Given that circle geometry is predominantly used for scatter points in scatter, dot, and bubble plots, our proposed method is generalized for our selected four chart types.

Thus, our contribution is integrating an end-to-end system, ACCirO (Analyzer for Chart images with C*ir*cular Objects), generalized for four chart types. ACCirO has a four-step workflow (Figure 1): $\mathbf{S_1}$ the chart classification, $\mathbf{S_2}$ a novel color-based annotation along with text extraction, $\mathbf{S_3}$ a novel color-based data extraction, $\mathbf{S_{4a}}$ text summarization and $\mathbf{S_{4b}}$ chart reconstruction. ACCirO specifically works for charts where color encodes class information and improves on $\mathbf{S_2}$ and $\mathbf{S_3}$ in BarChartAnalyzer [4], and ScatterPlotAnalyzer [5]. *Alpha blending* leads to the blended colors in bubble plots, which are different from the colors given in the chart legend. Our algorithm in $\mathbf{S_3}$ addresses the challenge of computing the color, radius, and center of constituent circles in overlap regions.

Related Work: We generalize data extraction for different chart types, as is the current focus [2, 4, 5, 8], but by using color information exclusively. CHT [6] has been used for object extraction from charts [8]. For the chart-types with circular objects in the foreground with non-textured background, CHT suffices.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_50

2 The Workflow of ACCirO

We propose a fully automated workflow for ACCiro for pie charts and dot, bubble, and scatter plots. We consider bubble and dot plots as variants of scatter plots, owing to the similarity in using circular objects as graphical objects with positional information. The four key components of our workflow (Figure 1) are: ($\mathbf{S_1}$) chart type classification, $\mathbf{S_2}$ chart component detection and extraction, and $\mathbf{S_3}$ data table extraction, optionally followed by $\mathbf{S_{4a}}$ chart summarization or $\mathbf{S_{4b}}$ chart reconstruction. We use the implementation from our previous work [4, 5] for $\mathbf{S_1}$ to pick scatter plots and pie charts, $\mathbf{S_{4a}}$, and $\mathbf{S_{4b}}$. Scatter plots are further subclassified to its variants based on position and size variations of scatter points.

S_2: Chart Component Detection & Extraction:
A chart is structurally composed of specific elements, referred to as *chart components*, whose characteristic properties in its raster format are exploited for their extraction. The seven components are: *canvas*, *legend*, *chart title*, *XY-axis titles*, and *XY-axis labels*. The region of the chart that *contains* the graphical objects, *e.g.,* pie sectors, scatter points, and bounded by axes, is the *canvas*. The process of localizing and retrieving them from the chart images is called *chart component extraction*. A separate component-wise analysis is more effective for stepwise chart interpretation than joint extraction using the entire image.

We first extract textual content by using a DL-based Optical Character Recognition (OCR), namely, Character Region Awareness for Text Detection (CRAFT), followed by a scene text recognition framework (STR), as used in ScatterPlotAnalyzer. This text is now removed to extract the canvas and graphical markers/objects in the legend. But, the filtered image still contains “noise” such as axis lines, gridlines, ticks, and small text fragments missed by the OCR.

In pie charts, we use CHT to extract the entire pie from the image, and the unique pixel colors in the pie pixels are used to locate objects in the legend. In scatter plots and their variants, object extraction using geometry is not reliable owing to the variety in marker styles and the presence of overlapping scatter points. So we initially remove axes and gridlines based on the property of the periodic arrangement of their straight pixels lines. We then locate the color-based clusters of pixels and tag them as “objects.” The color histogram of the image gives colors of high frequency needed for the localization of pixel clusters. The bounding box of pie and axes in scatter plots gives the canvas.

Legend extraction follows after the canvas extraction in our workflow to accommodate cases of legend being placed in the canvas. The color-based clusters with relatively smaller pixel coverage and placed adjacent to text are identified as legend markers with corresponding labels. The final step is to semantically classify the textboxes outside the canvas region based on their role, *i.e.,* chart title, axes titles, and labels. The axes are usually found at the bottom, and the left of the canvas region, and the chart title at its top. In legend-free pie charts, text boxes in the proximity of arc centers of the pie sectors give the class information.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_50

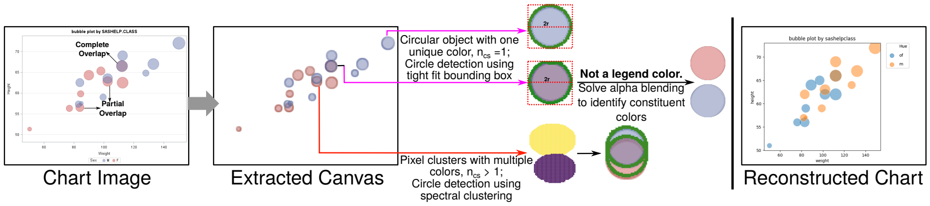

bubble plot by SASHELP.CLASS
Complete Overlap
Partial Overlap
Height
Weight
Sex M F
Chart Image
Extracted Canvas
Circular object with one unique color, n_{cs} =1; Circle detection using tight fit bounding box
2r
Not a legend color.
Solve alpha blending to identify constituent colors
Pixel clusters with multiple colors, n_{cs} > 1; Circle detection using spectral clustering
bubble plot by sashelpclass
Hue
of
m
height
weight
Reconstructed Chart

Fig. 2: Our proposed method for circular object extraction from bubble chart images.

S_3: Data Table Extraction from Graphical Objects:
We extract information from the pixel clusters in the image space and convert them to data space using the cues from the extracted text.

Pixel-based Data Extraction from Pie Charts: Percentage data is obtained from the sector area, which is determined using the fraction of pixel counts in the sector [2]. We implement this on the "donut" with one-third of the radius of the pie removed from the center. The donut is used instead of the entire pie to reduce the discrepancy from the missing pixels owing to text removal in the pie region. The sectors are then mapped to their labels based on colors. In legend-free charts, sectors are mapped to the closest text labels.

Pixel-based Data Extraction from Scatter, Dot, and Bubble Plots: Here, the 2D data is encoded in the positional vectors of the scatter object. The variants use additional visual encodings for more attributes. Such as, the height of stacks of dots in dot plots represents bar height in an equivalent bar chart, and bubble plots use the size and color of objects to encode additional attributes. Contours are extracted for color pixel clusters from]step2. The positional information of scatter points is determined using the contour centroids of the clusters. In the case of overlapping points, the number of points involved is computed based on the ratio of contour area with the smallest contour observed in the chart. We then use k-means clustering to get centroids in the overlapping region. In the case of a dot plot, we determine the count of objects stacked with the same x-coordinate value of contour centroids. This count is given a class label using the legend color or the x-tick mark label.

Bubble plots have the unique challenge of overlapping circular objects with varying sizes and transparencies. The resultant alpha blending of overlapped scatter object regions poses a challenge in identifying the number and parameters of the constituent scatter objects. The resultant color from alpha blending is given by: $C = \alpha.F + (1-\alpha).B$, where F and B are the foreground and background colors respectively; and $0 \leq \alpha \leq 1$. We resolve the challenge using contour colors (Figure 2). To estimate the number of overlapping points in a contour, we get n_{cs} segments based on the unique colors in contours. Using the value of n_{cs}, we extract the center, radius, and class label of the scatter point. If $n_{cs} = 1$, the center and radius of the scatter point are given by a tight-fitting bounding box of the contour. When $n_{cs} > 1$, we use spectral clustering in the pixel cluster of contour and determine circle parameters with best-fit circle regression of n_{cs}

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_50

clusters. In all cases, we use the class label corresponding to its contour color, as given in the legend.

For those circles which do not correspond to legend colors in the above two cases, we use the blending equation to solve for the legend colors and corresponding transparency, α, values that give the resultant blended colors C. We use an *exhaustive* search of the solution space to find the closest solution. We assume that the color C of the spectral cluster is a blend of two or more legend colors. Hence, we solve for $2 \leq p \leq m = |C_L|$, for m legend colors and the set of legend colors C_L. We use mC_p combinations of colors, where each experimental run uses a subset of legend colors $\{C_1, C_2, \ldots, C_p\}$, where $C_i \in C_L$, $\forall i \in [1, p]$. The blending equation is now modified as a linear combination of C_i:

$$C' = \sum_{i=1}^{p} \alpha_i . C_i, \text{ where } \sum_{i=1}^{p} \alpha_i = 1 \text{ and } 0 < \alpha_i < 1, \forall i \in [1, p].$$

We can reduce the search space by limiting the value of α_1 to be a value in the interval $[0.5, 0.95]$, with a step size of 0.05. We also implement a greedy algorithm terminating the search when the closest color is obtained.

Data Transformation for Scatter Plots: Finally, to transform data in pixel space to the numerical space, we use the scaling factor computed from the semantics of the text, as done in ScatterPlotAnalyzer. However, this does not work for circle size encoding in bubble charts, as the factor for circle sizes can be obtained from either the size legend (as present in some charts) or the parameter setting used in different plotting tools *e.g.,* DPI, for the chart generation. Thus, the exact data mapping for bubble charts will be explored in future.

3 Experiments and Results

Qualitative Assessment: Through the visualizations, we observe that our method performs superior to the tensor field-based method in ScatterPlotAnalyzer [5] (Figure 3). The color-based data extraction technique is an improvement over tensor fields [5] in the case of cluttered scatter points. Figure 4 shows outputs at different stages of ACCirO for a sample of each chart type, demonstrating similarities between the reconstructed and sources chart images. Visual analysis of the results of ACCiro on pie charts and scatter plots images in the FigureQA [7] gives reconstruction accuracy of ∼90.11% and ∼90.5%, respectively. To improve the circle detection using CHT [6] in a pie chart, which has an average of 96% accuracy, advanced circle detection methods such as RANSAC may be used.

Quantitative Assessment: For quantitative assessment, we use synthetically generated chart images from publicly available data sources, *e.g., Kaggle*. We have generated a set of 15 images each for pie charts, dot and bubble charts, and 24 for scatter plots. In total, we use a test set of 69 images here.

We use the F1 Score and MAPE (Mean Absolute Percentage Error) metrics to measure the success and failure rates of the performance of ACCirO (Table 1). We consider the data extraction as a *success* with F1 Score>0.8 as in [3]. Despite the smaller test dataset, the data extraction accuracy of ACCirO from scatter plots surpasses the state-of-the-art methods with F1-Score 97%, compared to

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_50

90.5% and 88% for MECDG [1] and Scatteract [3], respectively. Our data extraction for dot plots is 100% accurate, owing to the structured point layout. Even for bubble plots, despite the complex challenges due to transparency and overlapping points, we get an F1-Score of 100%. It must be noted that, since the bubble/object size is in pixel measure, we exclude it from the F1 score computation. We observe that the normalized radius values are closer to raw data.

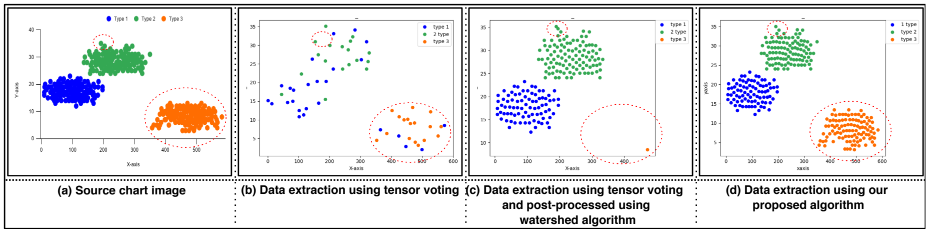
To determine the numerical precision errors in $\mathbf{S_3}$, we compare the difference between the source and extracted data values using the Mean Absolute Percentage Error (MAPE) as in [4]. Here, our alternative definition for the *success* of data extraction is when the error rate MAPE<0.2. MAPE is augmented in the case of omission and precision errors owing to cluster centroids overlapping with scatter points and pixel space to data space transformations, respectively. Pie charts have been an exception for error-free data extraction of percentage values.

4 Conclusions

ACCirO has two known limitations which are to be resolved in future work. Firstly, owing to ACCiro being a color-based method, it fails in cases where the shape and texture of scatter points encode class or type information. Secondly, the STR text recognition model fails to interpret superscript symbols, and recognition of 'o','0', and '−.' In summary, our proposed color-based end-to-end chart image interpretation system, ACCirO, has been generalized for the chart with circular objects, such as pie charts, and scatter, dot and bubble plots.

References

1. Chen, L., Zhao, K.: An Approach for Chart Description Generation in Cyber–Physical–Social System. Symmetry **13**(9), 1552 (2021)
2. Choi, J., Jung, S., Park, D.G., Choo, J., Elmqvist, N.: Visualizing for the non-visual: Enabling the visually impaired to use visualization. In: Computer Graphics Forum. vol. 38, pp. 249–260. Wiley Online Library (2019)
3. Cliche, M., Rosenberg, D., Madeka, D., Yee, C.: Scatteract: Automated extraction of data from scatter plots. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 135–150. Springer (2017)
4. Dadhich, K., Daggubati, S.C., Sreevalsan-Nair, J.: BarChartAnalyzer: Digitizing Images of Bar Charts. In: Intl. Conf. on Image Proc. and Vision Engg. (IMPROVE). pp. 17–28. INSTICC, SciTePress (2021)
5. Dadhich, K., Daggubati, S.C., Sreevalsan-Nair, J.: ScatterPlotAnalyzer: Digitizing Images of Charts Using Tensor-based Computational Model. In: Intl. Conf. on Computational Sc. – ICCS 2021, Part V, LNCS, vol. 12746. pp. 70–83 (2021)
6. Duda, R.O., Hart, P.E.: Use of the Hough transformation to detect lines and curves in pictures. Communications of the ACM **15**(1), 11–15 (1972)
7. Kahou, S.E., Atkinson, A., Michalski, V., Kádár, Á., Trischler, A., Bengio, Y.: FigureQA: An Annotated Figure Dataset for Visual Reasoning. CoRR **abs/1710.07300** (2017)
8. Savva, M., Kong, N., Chhajta, A., Fei-Fei, L., Agrawala, M., Heer, J.: Revision: Automated Classification, Analysis and Redesign of Chart Images. In: Proceedings of the 24th Annual ACM Symp. on User Interf. Softw. & Tech. pp. 393–402 (2011)



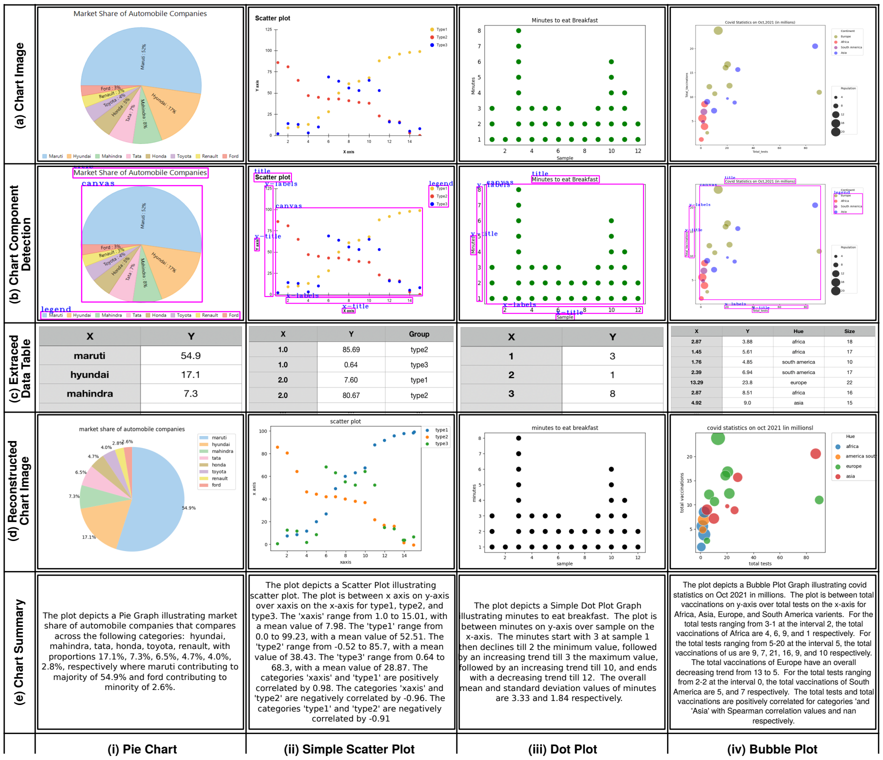
(a) Source chart image

(b) Data extraction using tensor voting

(c) Data extraction using tensor voting and post-processed using watershed algorithm

(d) Data extraction using our proposed algorithm

Fig. 3: Comparison of (a) source image and charts reconstructed using similar methods: (b) ScatterPlotAnalyzer [5], (c) modified ScatterPlotAnalyzer, and (d) our color-based method, for a multiclass scatter plot with a high degree of overlap of scatter points. The cluster of scatter points (red dotted ellipses) extracted is highlighted.



X	Y
maruti	54.9
hyundai	17.1
mahindra	7.3

X	Y	Group
1.0	85.69	type2
1.0	0.64	type3
2.0	7.60	type1
2.0	80.67	type2

X	Y
1	3
2	1
3	8

X	Y	Hue	Size
2.87	3.88	africa	18
1.45	5.61	africa	17
1.76	4.85	south america	10
2.39	6.94	south america	17
13.29	23.8	europe	22
2.87	8.51	africa	16
4.92	9.0	asia	15

The plot depicts a Pie Graph illustrating market share of automobile companies that compares across the following categories: hyundai, mahindra, tata, honda, toyota, renault, with proportions 17.1%, 7.3%, 6.5%, 4.7%, 4.0%, 2.8%, respectively where maruti contributing to majority of 54.9% and ford contributing to minority of 2.6%.

The plot depicts a Scatter Plot illustrating scatter plot. The plot is between x axis on y-axis over xaxis on the x-axis for type1, type2, and type3. The 'xaxis' range from 1.0 to 15.01, with a mean value of 7.98. The 'type1' range from 0.0 to 99.23, with a mean value of 52.51. The 'type2' range from -0.52 to 85.7, with a mean value of 38.43. The 'type3' range from 0.64 to 68.3, with a mean value of 28.87. The categories 'xaxis' and 'type1' are positively correlated by 0.98. The categories 'xaxis' and 'type2' are negatively correlated by -0.96. The categories 'type1' and 'type2' are negatively correlated by -0.91

The plot depicts a Simple Dot Plot Graph illustrating minutes to eat breakfast. The plot is between minutes on y-axis over sample on the x-axis. The minutes start with 3 at sample 1 then declines till 2 the minimum value, followed by an increasing trend till 3 the maximum value, followed by an increasing trend till 10, and ends with a decreasing trend till 12. The overall mean and standard deviation values of minutes are 3.33 and 1.84 respectively.

The plot depicts a Bubble Plot Graph illustrating covid statistics on Oct 2021 in millions. The plot is between total vaccinations on y-axis over total tests on the x-axis for Africa, Asia, Europe, and South America varients. For the total tests ranging from 3-1 at the interval 2, the total vaccinations of Africa are 4, 6, 9, and 1 respectively. For the total tests ranging from 5-20 at the interval 5, the total vaccinations of us are 9, 7, 21, 16, 9, and 10 respectively. The total vaccinations of Europe have an overall decreasing trend from 13 to 5. For the total tests ranging from 2-2 at the interval 0, the total vaccinations of South America are 5, and 7 respectively. The total tests and total vaccinations are positively correlated for categories 'and 'Asia' with Spearman correlation values and nan respectively.

(i) Pie Chart

(ii) Simple Scatter Plot

(iii) Dot Plot

(iv) Bubble Plot

Fig. 4: Different stages of chart data extraction followed by chart reconstruction and summarization from the sample source images of (i) pie chart, (ii) scatter plot, (iii) dot plot, and (iv) bubble plot. The text summary is best visible at 220+% zoom level.

Table 1: Accuracy of data table extraction using ACCirO

Chart Type → Accuracy Measure ↓	Pie Chart	Dot Plot	Bubble Plot	Scatter Plot	Overall Measures
Average Precision	0.94	1.00	1.00	0.97	**0.96**
Average Recall	1.00	1.00	0.99	0.95	**0.95**
Success Rate: F1 score > 0.8	93%	100%	100%	96%	**96.4%**
Success Rate: MAPE < 0.2	100%	100%	93%	84%	**88.7%**