transformations on the image. Most recent work by Savva et al.

2http://arohatgi.info/WebPlotDigitizer/app/

set of very complicated features, extracted through various

1http://academic.research.microsoft.com/

extracted from vector graphics. Prasad et al. [12] used a

graphs from our dataset to understand the speciﬁc challenges

Shao et al. [15] and Huang et al. [7] used low-level graphemes

We analyzed more than hundred randomly selected line

studied problem, and multiple features have been proposed.

Classiﬁcation of computer generated charts is a well-

and this work aims to bridge that gap.

not scalable. A fully automated system doesn’t exist till date,

II. RELATEDWORK

to regenerate that table, but manual methods

are tedious and

2

can not be accessed from the paper. It is extremely beneﬁcial

solves that problem.

ﬁgure 1, a line graph that was generated from a data table that

distinct RGB color values due to anti-aliasing. Our algorithm

methods and are generated from data tables. For example, see

distinguishable” colors, it might have more than one thousand

plotting region. These ﬁgures are used to compare multiple

While a color plot usually contains less than ten “visually

Line graphs are plots with single or multiple curves in the

process is trivial, and a color based segmentation would sufﬁce.

curves. As we consider color plots, it might appear that the

Preliminary analysis of the dataset is presented in section IV.

assigning each non-text pixel in the plot region to one of the

paper focuses on line graphs that are a subset of these ﬁgures.

a “data point”. The next step is the curve segmentation i.e.

extracted more than 40,000 ﬁgures from scholarly papers. This

properly, every pixel in the plot region can be mapped to

ﬁgures from scholarly PDF documents. Using their method we

once two X-axis values and two Y-axis values are identiﬁed

et al. [4] has proposed methods for automated extraction of

axes value/labels (see section VI-C). It is easy to see that

data (caption/mention) extraction. A recent paper by Clark

Next, extracted text is classiﬁed in seven classes such as

Previously, Ray Chaudhury et al. [3] explored ﬁgure meta-

in recent times. Our system improves on their method.

ﬁgures have received less attention [2].

the metrics standardized by document analysis community [8]

been many works about extraction and understanding of tables,

from plot images, but their method was not evaluated using

contained at least one ﬁgure and one table. While there have

Kataria et al. [9] proposed heuristics for text extraction

than 43% contained at least one table, and more than 36%

and 2014, more than 70% contained at least one ﬁgure, more

3) Extract the curves and associate them with the legends.

lished in top 50 computer science conferences

between 2004

1

values and Y-axis values are necessary and sufﬁcient.

For example, among 10,000 randomly selected articles pub-

Assuming the plot scales are linear, two pairs of X-axis

Scholarly papers usually contain many ﬁgures and tables.

region needs to be mapped into a (X-value,Y-value) pair.

I. INTRODUCTION

ends. For the data extraction, every point in the plot

2) Classify the words as X-axis/Y-axis value/labels or leg-

partially annotated dataset for future research is described.

1) Extract text “words” from the ﬁgure.

extraction method that has an average accuracy of 82%. A large

extraction is:

the previous method by 29%. We also propose a novel curve

the text extraction, our heuristics outperforms the accuracy of

For color line graphs, the generic algorithm for data

outperforms traditional low-level image descriptors by 10%. For

feature descriptors such as SIFT or HoG.

For the classiﬁcation, we show that unsupervised feature learning

show that unsupervised feature learning outperforms traditional

graphs; text extraction from the ﬁgures and curve extraction.

features for this problem has been proposed, we are the ﬁrst to

has multiple components: image classiﬁcation for identifying line

automated data extraction from color line graphs. Our system

(bar graphs, pie chart, photograph) instance. While multiple

automated data extraction system. Next, we describe a system for

an input image as a positive (color line graph) or negative

analysis of line graphs to explain the challenges of building a fully

The ﬁrst module in our system is a classiﬁer that classiﬁes

for such data extraction task is not yet available. We report an

with separate colors.

hard and not scalable. On the other hand, automated systems

ﬁgures can not be accessed. Manual extraction of this data is

cases: line graphs where the curves or data points are drawn

to compare performances of various methods. The data in these

cases for data extraction. Our current system focuses on easy

They are usually generated from a data table and often used

analysis (section III), we were able to identify easy and hard

Abstract—Line graphs are ubiquitous in scholarly papers.

of building a fully automated data extraction system. From our

sagnik@psu.edu

sxw327@psu.edu

pmitra@ist.psu.edu, giles@ist.psu.edu

Pennsylvania State University

Pennsylvania State University

Pennsylvania State University

Information Sciences and Technology

Computer Science and Engineering

Information Sciences and Technology

Sagnik Ray Choudhury

Shuting Wang

Prasenjit Mitra, C. Lee Giles

Graphs

Automated Data Extraction from Scholarly Line

difﬁcult? And 2. Can we identify easy and hard instances of

metadata contains following information:

in the plotting styles that make the text/graphics extraction

and a JSON metadata ﬁle for each ﬁgure in the document. The

Our research questions are: 1. What are the variations

“pdfﬁgures” [4]. Their system produces a grayscale image ﬁle

using a recently released system by Clark et al. known as

data.

ferences between 2004 and 2014. The ﬁgures were extracted

to integrate into an architecture due to the variability of the

10,000 articles published in top ﬁfty computer science con-

previous work has explored parts of the problem, they are hard

Our entire dataset consists of 40,000 ﬁgures extracted from

need to extract all these information from the ﬁgure. While

and mentions. An automated system for data extraction would

IV. DATASET

label is also a metadata for the graph, along with captions

texts are the column headers for these curves. The ﬁgure

inside the plot region as in the previous work.

multiple Y-values, corresponding to multiple curves. Legend

we do not make any assumption about the legend region being

headers for the columns in the table. For one X-value, there are

these problems. Also, for the text extraction and classiﬁcation,

the “data value” for the curve pixels. The axes labels are the

3. Overlapping curves. Section VI-C reports our methods for

table as shown in ﬁgure 1. Axes values are used to generate

ground, 2. Identifying “visually distinguishable colors” and

and 7. Other text. The metadata structure for a line graph is a

plots: 1. Removal of the grid structure and non-white back-

value 3. X-axis label, 4. Y-axis label, 5. Figure label, 6. Legend

There are three challenges in curve extraction from color

classiﬁed as one of the following: 1. X-axis value, 2. Y-axis

focus of this paper.

A line graph has curves and text words. A word can be

that have not been explored properly. These plots are the

III. ANALYSIS OFLINEGRAPHS FORDATAEXTRACTION

2) There are quite a few “easy” instances of the problem

in previous works.

plotting styles, and that aspect has gone largely unnoticed

aspect of the problem.

1) The hardness of the problem lies in the variation of the

graphs, but only for continuous curves. We explore another

architecture [11] reported curve extraction techniques for line

To summarize:

most relevant to ours. A more comprehensive version of their

is easier for these plots.

automated data extraction from 2D plots and their work is

these plots (89%) that is the case. Naturally, curve extraction

to all ﬁgures. Kataria et al. [9] proposed an architecture for

curve is plotted with a separate color. However, for most of

Therefore, it would be hard to generalize their approaches

Obviously, a color plot doesn’t automatically imply that each

vector graphics embedded in PDFs and not raster graphics.

by their color. 42% of plots in our dataset are color plots.

by Shao et al. [14], [15] these graphemes were extracted from

(see ﬁgure 2). 2.Color plotswhere curves can be distinguished

important step for the data extraction. In more recent works

connected. That assumption is not valid in most real graphs

were extracted from the images, which would be the most

for this problem, but they assumed that the curves were always

For example, they don’t clarify much how these graphemes

markers or other patterns. Liu et al. [11] proposed methods

focus much on the automated data extraction problem itself.

are plotted with black/gray pixels and are distinguished by

advanced the understanding of scholarly ﬁgures, they did not

in two classes: 1. Binary/ grayscale plots where curves

parsing the graphemes [6]. While these works have greatly

For curve extraction, line graphs can broadly be classiﬁed

Subsequent work proposed a more complete grammar for

pyramidal spatial index for efﬁcient computations of GERs.

inside the plotting region is not valid.

elements of a ﬁgure (symbols, lines, curves) and proposed a

algorithm, and 2. The assumption of legend regions being

“Generalized Equivalence Relations” (GER) between graphical

to be removed from the plotting region before applying any

Early papers by Futrelle [5] introduced the concept of

ods: 1. The “grid” structure and non-white background needs

These statistics indicate two problems with existing meth-

engineering. Our work reinforces that ﬁnding (section V).

[13] showed unsupervised feature learning outperforms feature

legend regions were not inside the plotting region.

nonideal plotting regions that had grid structures, as well as

that these characteristics were not exclusive, i.e. there were

between these classes and the data table is apparent.

and 4. Plotting region background was nonwhite (10%). Note

generated. It also shows the seven classes of text inside the ﬁgure. The mapping

the plotting region which were neither legend nor curve (15%)

Fig. 1. An example 2D line graph and the data table from which it was

plotting region (13%); 3. There were text/ graphic elements in

87%; 2. Legend region was not present (15%) or not in the

1. The plotting region had a grid structure (as in ﬁgure 2):

that there were four main reasons for plots being “nonideal”:

in our dataset had such an ideal plotting region. We observed

region. We ﬁnd that is often not the case. Only 58% of the plots

contains only two components: 1. Curves and 2. A legend

assumption is that the plotting region is always “ideal”, i.e.,

segmented into X-axis, Y-axis and plotting region. Another

styles. Previous work [11] assumes that a line graph can be

Many problems arise from the variations in the plotting

line graphs sampled from a large collection (section IV).

the problem? To answer them, we analyzed more than hundred

smoothing and subsampling of an image by a Gaussian kernel,

images. While their method achieved good accuracy (71%), it

(DoG) pyramid. A Gaussian pyramid is created by repeated

tional K-means approach to extract text regions from patent

image using scale-space extrema in a difference-of-Gaussians

of the text sparsity [9], [17]. Zhu et al. [17] used a convolu-

The SIFT feature descriptor tries to ﬁnd key points in an

Tesseract, but they have less accuracy on these images because

images. One choice is to use off-the-shelf OCR systems such as

model.

Text extraction in our case is easier than natural scene

and HoG. We used these descriptors in a bag of words (BoG)

has used common low-level image descriptors such as SIFT

A. Word Extraction

or a pie chart, etc. Surprisingly, none of the previous methods

lem, where images are classiﬁed as a line graph, a bar graph

the image and classify them to aid the data extraction process.

works have explored a similar multiclass classiﬁcation prob-

The goal of this module is to extract the text words from

an input image is classiﬁed as a line graph or not. Previous

The ﬁrst step in our architecture is a classiﬁcation problem:

VI. WORDEXTRACTION ANDCLASSIFICATION

V. CLASSIFICATION OFFIGURES

publicly available.

HOG ANDSIFT.

LEARNING OUTPERFORMS LOW-LEVEL IMAGE DESCRIPTORS SUCH AS

for the problem analysis (section III). This dataset will be made

CLASSIFICATION: RANDOM PATCH BASED UNSUPERVISED FEATURE

design, a completely separate sample of 120 ﬁgures were used

TABLE I.

ACCURACY RESULTS FOR COMPUTER GENERATED CHARTS

discussed in section VI-B). To avoid bias in the algorithm

Average

80%

68%

40%

Venn Plot

75%

68%

42%

these ﬁgures were manually tagged with word class labels as

Table

86%

72%

46%

(to maintain consistency with previous work). A hundred of

Bar Plot

78%

66%

39%

in this paper except for the ﬁgure classiﬁcation experiment

Scatter Plot

79%

63%

40%

images were subsequently sampled for various experiments

Radar Plot

88%

70%

44%

Pie Plot

79%

71%

42%

a sub-ﬁgure. Among them, 882 were color plots. These 882

Pareto Plot

85%

67%

51%

either a line graph or contained at least one line graph as

Maps

84%

68%

52%

of 10,000 ﬁgures, we manually selected 2250 plots that were

Curve Plot

73%

64%

48%

Random (Savva et al. [13])

HoG

SIFT

line graphs, bar charts, and pie charts. From a random selection

of these ﬁgures contained sub-ﬁgures: often a combination of

sufﬁciently good representations.

We manually examined the images and found around 50%

contain complex structures and randomly selected patches are

document where the ﬁgure was mentioned [3].

(80%). This indicates that the scholarly charts usually do not

2. We modiﬁed the metadata to include the paragraphs in the

neither of them outperformed the random patches method

PDF as a grayscale image, therefore, were extracted as such.

(68%) performed better than the SIFT descriptor (40%) but

around 60% of these images were originally embedded in the

accuracy (mean of class speciﬁc accuracies), HoG descriptor

metadata, we re-extracted the ﬁgure as a color image. However,

data. The results are shown in table I. In terms of average

We made two modiﬁcations to the existing system: 1. From the

fold cross validation and compared the accuracies on the test

14.2%) and Venn diagrams (88 7.8%). We used stratiﬁed k-

are embedded as vector graphics (eps/ps/PDF).

charts (118 10.4%), radar plots (86 7.6%), scatter plots (159

(horizontal/vertical). Typically, this happens when ﬁgures

13%), maps (200 17.7%), Pareto charts (117, 10.3%), pie

The bounding box of the word, and 3. The orientation

in eight categories: bar charts (215, 19%), line graphs (147

ing metadata is available: 1. The text of the word, 2.

sake of comparison. In summary, we had 1130 images divided

PDF itself, then for each text word in the ﬁgure follow-

in natural image classiﬁcation. We used their dataset for the

 If the text inside the ﬁgure can be extracted from the

unsupervised feature learning that has shown excellent promise

 Caption of the ﬁgure.

descriptors. Their method is motivated by recent advances in

 Location of the ﬁgure on the page (bounding box).

random patches of pixels in a BoG model instead of image

 Page number for the page where the ﬁgure appears.

For the same classiﬁcation problem, Savva et al. [13] used

tion is used as a feature vector.

be separated by their patterns.

frequency distribution over these visual words. This distribu-

Fig. 2. A monochrome plot (extracted from [10]) where the curves can only

dictionary or codebook. Then, each image is represented as a

from the image and then clustered to create a “visual words”

interest points and descriptors for these points are extracted

The BoG method in our case works in two steps. First,

bar graphs and curves in the line graphs in the feature space.

them was to represent speciﬁc structures such as bars in the

detection and image classiﬁcation. Our motivation for using

Both HoG and SIFT have been extensively used in object

image and combines them to produce a ﬁnal feature descriptor.

occurrences of gradient orientation in localized portions of an

Histogram of Oriented Gradients (HOG) algorithm counts

ence between the adjacent levels in the Gaussian pyramid.

and the DoG pyramid is created by computing the differ-

text extraction from images. For each image, ground truth G

axis label. This is similar to pixel labeling problems in image

been used extensively in other works and competitions [8] for

label”, nearby words have a higher probability of being an X-

evaluation protocol is adopted from Wolf et al. [16] that has

neighbors. For example, if a word is classiﬁed as an “X-axis

were available, and we used them as the gold standard. Our

having a particular class label is dependent on the labels of its

ﬁgures being vector graphics, bounding boxes of the words

it is hard to train a classiﬁer. However, the probability of a text

that were originally embedded as vector graphics. These

word and 2. The location of the word on the image. Therefore,

IV) of extracted ﬁgures, we selected 200 color line graphs

We only have two features for each word: 1. The text of the

1) Experiments and Results: From our dataset (see section

largely unnoticed in the previous works.

essential in an automated data extraction process, it has gone

terminated in two or three iterations.

line graph as shown in ﬁgure 1. While this classiﬁcation step is

run iteratively until there is no merge. The process is usually

other words can be mapped to the metadata structure for the

heuristic worked best among them. The merging process is

X-axis values and two Y-axis values sufﬁce for that step. Also,

mean of the horizontal/vertical distances. The second minimum

to map each point in the plotting region to a data value.Two

thresholds such as minimum, second minimum, median and

are classiﬁed as “other text”. For the data extraction, we need

distance threshold. We experimented with several distance

that are used to show speciﬁc points of interest. These words

threshold. Other CCs are merged horizontally using a separate

of the ﬁrst six classes, sometimes line graphs contain words

vertically if the vertical distances between them are less than a

Other text. While most of the words can be classiﬁed as one

width from left are the candidates for Y-axis labels and merged

axis label, 4. Y-axis label, 5. Figure label, 6. Legend and 7.

in the previous work. CCs that are within 10% of the image

lowing seven classes: 1. X-axis value, 2. Y-axis value 3. X-

these two distances separately rather than combining them as

Extracted words need to be classiﬁed in one of the fol-

except for Y-axis labels. Therefore, it was beneﬁcial to use

 We observed that most texts in these ﬁgures were horizontal,

B. Word Classiﬁcation

characters. In the next step, they were merged to form words.

The output from this step was a set of CCs that are the text

baseline by 32%, 18%, and 29%.

an algorithm for that. In future, trained classiﬁers can be used.

are 63%, 75%, and 67% respectively, which is better than the

image because of anti-aliasing. In section VI-C we propose

to the poor result. Our average precision, recall, and F1 values

it is hard to identify black, gray or white pixels from an

to be wrong for most of our images. We believe that attributed

black pixels. Therefore, color based ﬁlter sufﬁced. However,

lines are considered as axes lines. We found that assumption

text in the color line graphs is almost always written using

tic, i.e. the longest pair of intersecting vertical and horizontal

 Candidate CCs were ﬁltered to remove noise. We found that

extraction depends on this step, but the segmentation is heuris-

candidates for the text characters.

(area below X-axis), Y-axis region and plot region. The text

 CCs having area<1% of the image area were considered as

The baseline method segments the image into X-axis region

We made following changes to the existing algorithm:

images, with our method and Kataria et al. [9] as baseline.

OB

OB

P

, R

and F1 values for the text extraction from 200

them.

Figures 3a, 3b and 3c show the comparative histograms of

from images. The previous algorithm was not evaluated using

proposed standardized evaluation metrics for text extraction

cases.

inputs. In recent times, the document analysis community has

many-to-one and one-to-many matchings are also valid in those

algorithm was non-iterative and depended on the order of

practice for these values are 0.4 and 0.8 [8] respectively, but

R

P

that needed further segmentation. Also, the proposed merging

The thresholds 

and 

are both set to 0.4. The standard

P

G

G

generated large text regions often containing multiple words

the ratio of the overlap between b

, b

and the area of b

.

P

G

P

step was needed to remove noise. The merging heuristic

tween b

, b

and the area of b

. Area-Recall is deﬁned as

area <1% were not text characters. Therefore, a ﬁltering

Area-Precision is deﬁned as the ratio of the overlap be-

area smaller than 1% of the image area. Also, all CCs having

jGj

OB

 R

=

found several problems with it. Most text characters had an

no. of correctly identiﬁed rectangles

While the CC approach was computationally efﬁcient, we

jPj

OB

 P

=

no. of correctly identiﬁed rectangles

training data.

 no. of correctly identiﬁed rectangles+=1

standard deviations of vertical and horizontal distances on the

P

G

R

1

2

Recall(b

,b

)>=

respectively. The parameters a,b,s

and s

are the mean and

P

G

P

i

j

 if Area-Precision(b

,b

) >=



and Area-

are the vertical and horizontal distances between C

and C

,

P

G

i

j

ij

ij

 for all pairs (b

,b

)

where P(C

; C

and C

ij

1

ij

2

x

:e(C

. Cy

x

b)=2s2

) = e(Cy

a)=2s2

i

j

i

j

OB

CCs C

and C

are considered “close” if P(C

; C

) > 0:05

and recall (R

) scores are calculated as follows:

OB

merged to form “words” if they are “close” to each other. Two

are not helpful for us. For a pair (P,G), the precision (P

)

in scholarly ﬁgures are larger than that. The rest of the CCs are

multiple boxes in the gold standard). These other matchings

than 20% of the image are discarded because rarely characters

single box) and one-to-many matching (one box predicted for

for the text characters. Therefore, the ones with an area greater

matching such as many-to-one (multiple boxes predicted for a

component (CC)s are extracted. These CCs are the candidates

to-one matching. Note that Wolf et al. allows for other types of

ﬁrst an edge map of the image is created and connected

overlap is deﬁned as the equivalence. This is deﬁned as one-

G

P

approach based on connected components. In their method,

For each box b

in G, the box in P (b

) having maximum

scale sliding windows. Kataria et al. [9] used a much simpler

predicted result P contains coordinates of j bounding boxes.

was computationally very expensive because they used multi-

contains coordinates of i bounding boxes for the words. A

the merged region. This process is continued until there is no

two RGB values doesn’t imply that the colors are visually

other, they are merged, and the bounding box is updated with

representation of colors as a small Euclidean distance between

labeled as “other text” are selected. If they are close to each

plots are extracted from PDFs. RGB space is not a “natural”

“legend box” which is initially empty. Iteratively the words

thousand distinct RGB values due to “anti-aliasing”, as these

the “other text” words as legends. A region is deﬁned as the

very few “visually distinguishable” colors, it has more than

we employed a region growing algorithm to label some of

plots with the legend text. While a color plot usually has

region of the ﬁgure. That motivated our third step, where

that separates out the curves in the plots and matches the

Usually, the legend region is the most textually dense

The ﬁnal module in our system is a heuristic algorithm

less than 5% of the image width or image height, respectively.

C. Curve Segmentation and Assignment

another, if the horizontal or vertical distance between them is

text” was close to it. A word is considered to be “close” to

classes was relabeled as “other text”, if a word of class “other

CLASSIFICATION OF WORDS IN THE FIGURES.

TABLE III.

PRECISION, RECALL ANDF1-SCORES FOR THE

seconds step, each word that was not classiﬁed as one of these

Other text

48%

71%

57%

value, Y-axis value and Y-axis label (see table III). In the

Legend

90%

93%

91%

were very accurate for the following three classes: X-axis

Figure label

83%

73%

78%

Y-axis label

95%

97%

96%

in the table II or “other text”. We observed that the heuristics

Y-axis value

97%

95%

95%

In the ﬁrst step, each text is classiﬁed as one of the classes

X-axis label

82%

95%

88%

X-axis value

99%

90%

95%

heuristics.

precision

recall

f1-score

classes such as legend or “other text” don’t have intuitive

heuristics for classes such as axes values and axes labels,

be considered as a conﬁdence point. While it is easy to design

CLASSIFICATION OF WORDS IN THE FIGURES.

TABLE II.

PRECISION, RECALL ANDF1-SCORES FOR THE

are summarized in table II. Each heuristics in the table II can

bel

nothing above the text.

We designed heuristics for the class prediction problem that

Figure la-

1. Within 10% of the image height from the top, 2. There’s

bel

the left of the text.

neighboring labels.

Y-axis la-

1. Text is vertical, 2. Not a number, 3. There’s nothing to

is that the labels can change iteratively depending on the

value

the left of the image.

Y-axis

1. Text is a number, 2. Within 10% from image width from

interaction between neighboring words. The key observation

bel

[1], but we developed a much simpler algorithm to model the

X-axis la-

2. Text is not a number, 2. There’s nothing below the text.

approximations based on max-ﬂow/min-cut techniques exist

value

the bottom of the image.

X-axis

1. Text is a number, 2. Within 10% of image height from

more than two labels, and we have seven labels. Efﬁcient

Class

Heuristics for classiﬁcation

instead of pixels. The minimization problem is NP-hard for

We are trying to predict the class label for the words

do not follow a logical layout structure.

this is because they are randomly spread over the images and

all pairs of neighboring pixels.

classiﬁed correctly, except for the class “other text”. We believe

q

L

(also known as “n interaction potential”). N is the set of

accuracy, without desigining any heuristic. Most words are

p

between two neighboring pixels p and q having labelsL

and

for two classes “legends” and “other text” with 91% and 57%

p;q

as a “data penalty function”), andV

models the interaction

III. It is important to note that we predicted the class labels

p

a cost function for the pixel p having the labelL

(also known

versus all precision, recall and f1 scores are reported in table

p

p

p

HereL=fL

jp2Pgis a labeling of the image P,D

(L

)is

from 100 ﬁgures in the aforementioned seven classes. One

p2P

p;q2N

For the experiment, we manually tagged the 2000 words

p

p

p;q

p

q

E(L) =

D

(L

) +

V

(L

; L

):

(1)

“other text”.

X

X

is minimized:

classiﬁed as legends and rest of the words are classiﬁed as

processing [1], where an energy equation such as equation 1)

merge possible. Finally, the words inside the “legend box” are

Fig. 3. Comprative results for text extraction: our method and Kataria et al. [9]

(a)

(b)

(c)

graphs.

accuracy and consider curve extraction from BW/grayscale line

International Conference on. IEEE, 2013, pp. 633–637.

papers. In future, we plan to improve the text extraction

boost,” in Document Analysis and Recognition (ICDAR), 2013 12th

brieﬂy discuss our dataset of ﬁgures extracted from scholarly

images using convolutional k-means feature quantization and ada-

classiﬁcation and curve extraction from line graphs. We also

[17] S. Zhu and R. Zanibbi, “Label detection and recognition for uspto

hard cases. Next, we present algorithms for text extraction,

296, 2006.

We present an analysis of line graphs to identify easy and

of Document Analysis and Recognition (IJDAR), vol. 8, no. 4, pp. 280–

of object detection and segmentation algorithms,”International Journal

architecture for automated data extraction for such line graphs.

[16] C. Wolf and J.-M. Jolion, “Object count/area graphs for the evaluation

important, can not be accessed or searched on. We report an

Perspectives. Springer, 2006, pp. 231–242.

are generated from tabular data and that data, while very

pdf documents,” inGraphics Recognition. Ten Years Review and Future

Line graphs are abundant in scholarly papers. These ﬁgures

[15] M. Shao and R. P. Futrelle, “Recognition and classiﬁcation of ﬁgures in

Proc. of GREC, 2005.

VII. CONCLUSION

[14] M. Shao and R. Futrelle, “Graphics recognition in pdf documents,” in

interface software and technology. ACM, 2011, pp. 393–402.

images,” in Proceedings of the 24th annual ACM symposium on User

left or right to the legend text.

“Revision: Automated classiﬁcation, analysis and redesign of chart

[13] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer,

We observed that for almost all plots, the color symbol appears

CBMI’07. International Workshop on. IEEE, 2007, pp. 85–92.

classiﬁed, the curve-legend assignment is easy for color curves.

puter generated charts,” in Content-Based Multimedia Indexing, 2007.

Once the curves are segmented and the text from the ﬁgure is

[12] V. Prasad, B. Siddiquie, J. Golbeck, and L. Davis, “Classifying com-

correctly predicted the number of colors in 82% of cases.

search,”IJDAR, vol. 12, no. 2, pp. 65–81, 2009.

colors. On a random sample of 165 images, our algorithm

“Automated analysis of images in documents for intelligent document

and the reconstructed curves and predicted the number of

[11] X. Lu, S. Kataria, W. J. Brouwer, J. Z. Wang, P. Mitra, and C. L. Giles,

For evaluation, we manually examined the original image

design & implementation. USENIX Association, 2007, pp. 22–22.

in Proceedings of the 4th USENIX conference on Networked systems

used.

[10] J. Ledlie, P. Gardner, and M. Seltzer, “Network coordinates in the wild,”

sequential error correction model common in OCR can be

intelligence, vol. 2, 2008, pp. 1169–1174.

even overlapping ones can be extracted easily. In future, a

documents,” inProceedings of the 23rd national conference on Artiﬁcial

of data points and text blocks from 2-dimensional plots in digital

drawn by solid bold lines. But for curves with dotted lines,

[9] S. Kataria, W. Browuer, P. Mitra, and C. Giles, “Automatic extraction

Overlapping curves pose a challenge when the curves are

Conference on. IEEE, 2013, pp. 1484–1493.

Document Analysis and Recognition (ICDAR), 2013 12th International

variable.

and L.-P. de las Heras, “Icdar 2013 robust reading competition,” in

heuristic is motivated by the idea of the rate of change in a

S. Robles Mestre, J. Mas, D. Fernandez Mota, J. Almazan Almazan,

1) - frequency of bin (i) > 0.5\* frequency of bin (i-1). This

[8] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez i Bigorda,

lower frequency: from 1 to i such that frequency of bin (i-

tives. Springer, 2004, pp. 87–99.

to the curves. Bins are selected iteratively from higher to

recognition,” in Graphics Recognition. Recent Advances and Perspec-

[7] W. Huang, C. L. Tan, and W. K. Leow, “Model-based chart image

each bin). The bins with high-frequency values correspond

IEEE, 1995, pp. 782–790.

the color bins based on their frequency (number of pixels in

1995., Proceedings of the Third International Conference on, vol. 2.

reduces to determining the actual number of colors. We sorted

using constraint-based parsing,” inDocument Analysis and Recognition,

Given a set of quantized “color” pixels, the problem

[6] R. P. Futrelle and N. Nikolakis, “Efﬁcient analysis of complex diagrams

of the 10th ICPR, 1990, pp. 403–408.

In future, we plan to investigate this problem further.

structures, animate vision, and generalized equivalence,” inProceedings

our algorithm. We sacriﬁce the recall in favor of precision.

[5] R. P. Futrelle, “Strategies for diagram understanding: Object/spatial data

drawn with black pixels, and hence would be removed by

tables, and captions from computer science paper,” 2015.

This approach solved that problem. Indeed, some curves are

[4] C. Clark and S. Divvala, “Looking beyond text: Extracting ﬁgures,

is the presence of the grid structure in the plotting region.

Conference on. IEEE, 2013, pp. 135–139.

in section III we mentioned that one problem with color graphs

Document Analysis and Recognition (ICDAR), 2013 12th International

and C. L. Giles, “Figure metadata extraction from digital documents,” in

classiﬁed as white,black or gray, and hence removed. Note that

[3] S. R. Choudhury, P. Mitra, A. Kirk, S. Szep, D. Pellegrino, S. Jones,

saturation parameter is less than eleven, that pixel could be

information retrieval. ACM, 2006, pp. 581–588.

suggested that if the “value” parameter is less than six and the

international ACM SIGIR conference on Research and development in

However, determining such threshold is hard. Our experiments

resource for digital libraries,” in Proceedings of the 29th annual

can be removed. These pixels have low saturation or value.

[2] S. Carberry, S. Elzer, and S. Demir, “Information graphics: an untapped

Therefore, it was evident that white, gray and black pixels

pp. 1124–1137, 2004.

color plots very rarely curves are drawn with black pixels.

Analysis and Machine Intelligence, IEEE Transactions on, vol. 26, no. 9,

cut/max-ﬂow algorithms for energy minimization in vision,” Pattern

gray pixels are almost always background values and 3. In

[1] Y. Boykov and V. Kolmogorov, “An experimental comparison of min-

are almost always written in black colors, 2. White and

Three important observations are: 1. Text in the plots

REFERENCES

max 360).

Foundation).

bins based on their hue value (bin size=10, min hue value 0,

from the Qatar National Research Fund (a member of Qatar

the color. Therefore, we ﬁrst quantized the pixels in 36 color

tional Science Foundation and NPRP grant # 4-029-1-007

the saturation and value components determine the shade of

We gratefully acknowledge partial support from the Na-

space. In HSV, the hue component determines the color and

close. Therefore, we converted the images into HSV color

VIII. ACKNOWLEDGMENTS