Article

DeHyFoNet: Deformable Hybrid Network for Formula Detection in Scanned Document Images

Muhammad Zeshan Afzal [1,2,3,†*], Khurram Azeem Hashmi [1,2,3,†], Alain Pagani [3], Marcus Liwicki [4] and Didier Stricker [1,3]

[1] Department of Computer Science, Technical University of Kaiserslautern, 67663 Kaiserslautern, Germany; khurram_azeem.hashmi@dfki.de (K.A.H.); didier.stricker@dfki.de (D.S.)
[2] Mindgarage, Technical University of Kaiserslautern, 67663 Kaiserslautern, Germany
[3] German Research Institute for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany; alain.pagani@dfki.de
[4] Department of Computer Science, Luleå University of Technology, 971 87 Luleå, Sweden; marcus.liwicki@ltu.se
* Correspondence: muhammad_zeshan.afzal@dfki.de
† These authors contributed equally to this work.

Abstract: This work presents an approach for detecting mathematical formulas in scanned document images. The proposed approach is end-to-end trainable. Since many OCR engines cannot reliably work with the formulas, it is essential to isolate them to obtain the clean text for information extraction from the document. Our proposed pipeline comprises a hybrid task cascade network with deformable convolutions and a Resnext101 backbone. Both of these modifications help in better detection. We evaluate the proposed approaches on the ICDAR-2017 POD and Marmot datasets and achieve an overall accuracy of 96% for the ICDAR-2017 POD dataset. We achieve an overall reduction of error of 13%. Furthermore, the results on Marmot datasets are improved for the isolated and embedded formulas. We achieved an accuracy of 98.78% for the isolated formula and 90.21% overall accuracy for embedded formulas. Consequently, it results in an error reduction rate of 43% for isolated and 17.9% for embedded formulas.

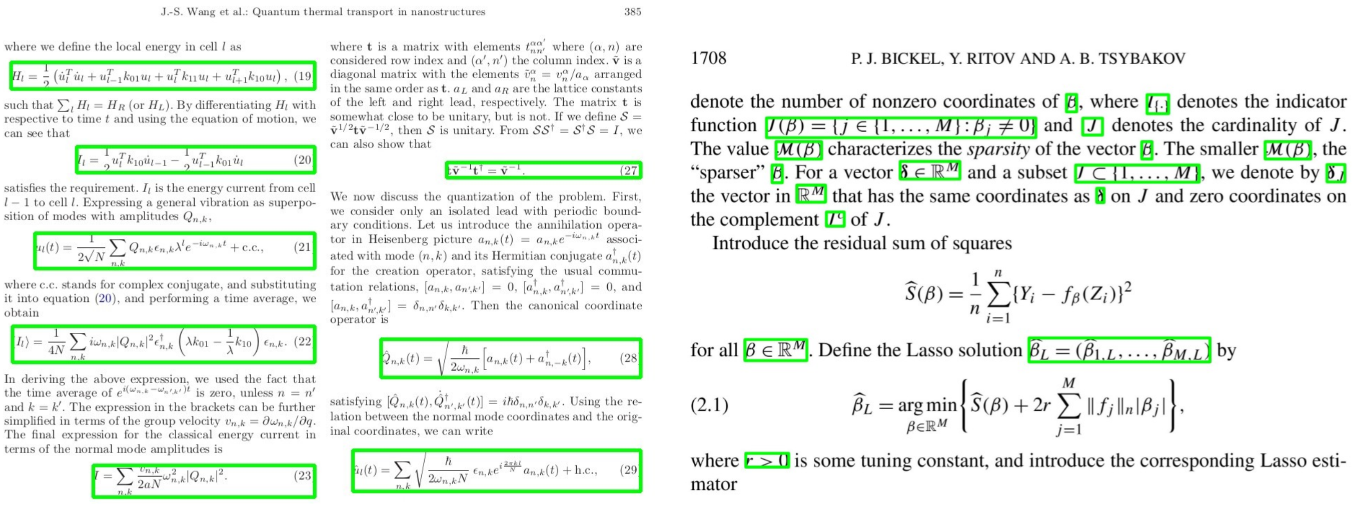
Keywords: formula detection; Hybrid Task Cascade network; mathematical expression detection; document image analysis; deep neural networks; computer vision

1. Introduction

Mathematical formulas are universally used to explain complex information compactly. Therefore, reliable detection of formulas is a primary step in digitizing scientific scanned documents. Mainly, the formulas are further categorized into two categories which are isolated formulas and embedded formulas [1]. Like other page objects, the isolated formulas are mentioned on a separate line, whereas the embedded formulas representing mathematical symbols are a part of a regular text line. Figure 1 demonstrates the detection of isolated and embedded formulas in a document image.

Several challenges involved in detecting isolated and embedded formulas include low inter-class variance (less dissimilarity between formulas and tables), high inter-class variance (more dissimilarity between two formulas), complex layouts, textual noise in documents, and diversity in mathematical symbols. Figure 1 exhibits a few instances of embedded formulas, which can be any mathematical notation written in a greek letter, mathematical operators, functions, and English alphabets declared as variables.

Earlier approaches tried to tackle this problem by exploiting Optical Character Recognition (OCR) systems [2,3]. Even after heavily relying on custom heuristics, most OCR-based systems produced an unsatisfactory interpretation of mathematical formulas in documents. The recent state-of-the-art OCR systems [3–6]. have gained huge success in interpreting textual content. However, there is still a massive gap in understanding



J.-S. Wang et al.: Quantum thermal transport in nanostructures 385

where we define the local energy in cell l as

$$H_l = \frac{1}{2}\left(\dot{u}_l^T \dot{u}_l + u_{l-1}^T k_{01} u_l + u_l^T k_{11} u_l + u_{l+1}^T k_{10} u_l\right), \quad (19)$$

such that $\sum_l H_l = H_R$ (or H_L). By differentiating H_l with respective to time t and using the equation of motion, we can see that

$$I_l = \frac{1}{2} u_l^T k_{10} \dot{u}_{l-1} - \frac{1}{2} u_{l-1}^T k_{01} \dot{u}_l \quad (20)$$

satisfies the requirement. I_l is the energy current from cell $l-1$ to cell l. Expressing a general vibration as superposition of modes with amplitudes $Q_{n,k}$,

$$u_l(t) = \frac{1}{2\sqrt{N}} \sum_{n,k} Q_{n,k} \epsilon_{n,k} \lambda^l e^{-i\omega_{n,k}t} + \text{c.c.}, \quad (21)$$

where c.c. stands for complex conjugate, and substituting it into equation (20), and performing a time average, we obtain

$$\langle I_l \rangle = \frac{1}{4N} \sum_{n,k} i\omega_{n,k} |Q_{n,k}|^2 \epsilon_{n,k}^{\dagger} \left(\lambda k_{01} - \frac{1}{\lambda} k_{10}\right) \epsilon_{n,k}. \quad (22)$$

In deriving the above expression, we used the fact that the time average of $e^{i(\omega_{n,k}-\omega_{n',k'})t}$ is zero, unless $n = n'$ and $k = k'$. The expression in the brackets can be further simplified in terms of the group velocity $v_{n,k} = \partial\omega_{n,k}/\partial q$. The final expression for the classical energy current in terms of the normal mode amplitudes is

$$I = \sum_{n,k} \frac{v_{n,k}}{2aN} \omega_{n,k}^2 |Q_{n,k}|^2. \quad (23)$$

where $\mathbf{t}$ is a matrix with elements $t_{nn'}^{\alpha\alpha'}$ where (α, n) are considered row index and (α', n') the column index. $\tilde{\mathbf{v}}$ is a diagonal matrix with the elements $\tilde{v}_n^{\alpha} = v_n^{\alpha}/a_{\alpha}$ arranged in the same order as $\mathbf{t}$. a_L and a_R are the lattice constants of the left and right lead, respectively. The matrix $\mathbf{t}$ is somewhat close to be unitary, but is not. If we define $\mathcal{S} = \tilde{\mathbf{v}}^{1/2}\mathbf{t}\tilde{\mathbf{v}}^{-1/2}$, then $\mathcal{S}$ is unitary. From $\mathcal{S}\mathcal{S}^{\dagger} = \mathcal{S}^{\dagger}\mathcal{S} = I$, we can also show that

$$\mathbf{t}\tilde{\mathbf{v}}^{-1}\mathbf{t}^{\dagger} = \tilde{\mathbf{v}}^{-1}. \quad (27)$$

We now discuss the quantization of the problem. First, we consider only an isolated lead with periodic boundary conditions. Let us introduce the annihilation operator in Heisenberg picture $a_{n,k}(t) = a_{n,k} e^{-i\omega_{n,k}t}$ associated with mode (n,k) and its Hermitian conjugate $a_{n,k}^{\dagger}(t)$ for the creation operator, satisfying the usual commutation relations, $[a_{n,k}, a_{n',k'}] = 0$, $[a_{n,k}^{\dagger}, a_{n',k'}^{\dagger}] = 0$, and $[a_{n,k}, a_{n',k'}^{\dagger}] = \delta_{n,n'}\delta_{k,k'}$. Then the canonical coordinate operator is

$$\hat{Q}_{n,k}(t) = \sqrt{\frac{\hbar}{2\omega_{n,k}}} \left[a_{n,k}(t) + a_{n,-k}^{\dagger}(t)\right], \quad (28)$$

satisfying $[\hat{Q}_{n,k}(t), \dot{\hat{Q}}_{n',k'}^{\dagger}(t)] = i\hbar\delta_{n,n'}\delta_{k,k'}$. Using the relation between the normal mode coordinates and the original coordinates, we can write

$$\hat{u}_l(t) = \sum_{n,k} \sqrt{\frac{\hbar}{2\omega_{n,k}N}} \epsilon_{n,k} e^{i\frac{2\pi k l}{N}} a_{n,k}(t) + \text{h.c.}, \quad (29)$$

(a) Isolated formulas

1708 P. J. BICKEL, Y. RITOV AND A. B. TSYBAKOV

denote the number of nonzero coordinates of β, where $I_{\{\cdot\}}$ denotes the indicator function $J(\beta) = \{j \in \{1, \ldots, M\} : \beta_j \neq 0\}$ and $|J|$ denotes the cardinality of J. The value $\mathcal{M}(\beta)$ characterizes the *sparsity* of the vector β. The smaller $\mathcal{M}(\beta)$, the "sparser" β. For a vector $\delta \in \mathbb{R}^M$ and a subset $J \subset \{1, \ldots, M\}$, we denote by δ_J the vector in $\mathbb{R}^M$ that has the same coordinates as δ on J and zero coordinates on the complement J^c of J.

Introduce the residual sum of squares

$$\widehat{S}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - f_\beta(Z_i)\}^2$$

for all $\beta \in \mathbb{R}^M$. Define the Lasso solution $\widehat{\beta}_L = (\widehat{\beta}_{1,L}, \ldots, \widehat{\beta}_{M,L})$ by

$$\widehat{\beta}_L = \underset{\beta \in \mathbb{R}^M}{\arg\min}\left\{\widehat{S}(\beta) + 2r\sum_{j=1}^{M}\|f_j\|_n|\beta_j|\right\}, \quad (2.1)$$

where $r > 0$ is some tuning constant, and introduce the corresponding Lasso estimator

(b) Embedded formulas

Figure 1. Figure 1a and Figure 1b highlight boundaries of isolated and embedded formulas, respectively. For brevity, separate images are used. The isolated formulas marked in Figure 1a are misclassified with other graphical page object such as figures and tables, whereas the embedding formulas emphasized in Figure 1b are mistreated as a regular textural content.

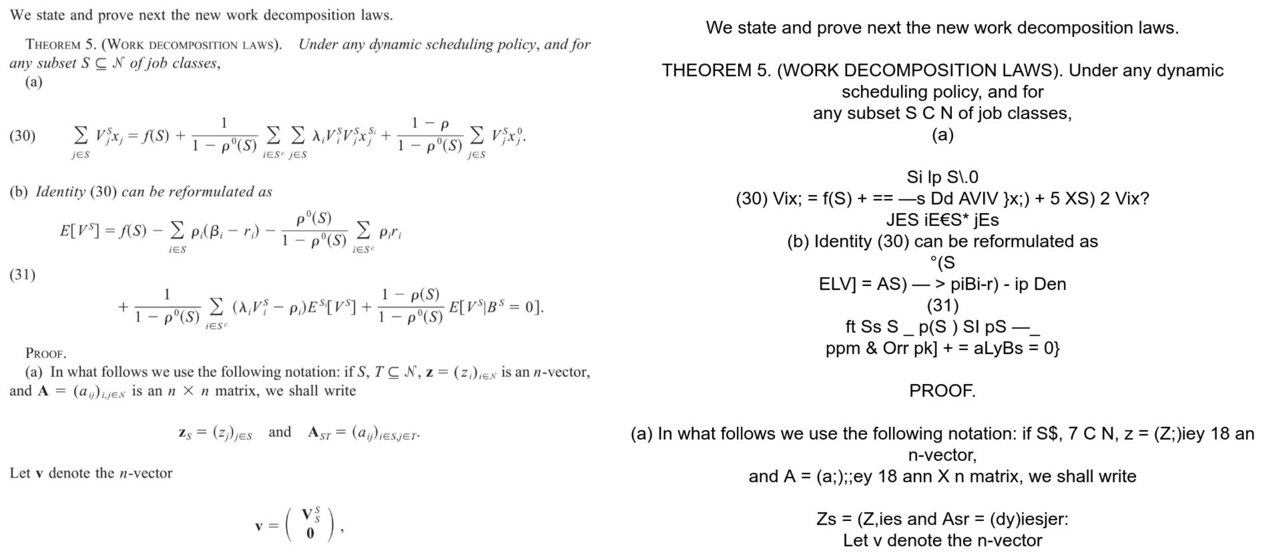
mathematical content in document images. To comprehend better, Figure 2 visualizes the huge difference between the understanding of textual content and mathematical content in scanned document images by an OCR. To generate Figure 2, we apply open-source LSTM based OCR, Tesseract [6] (available at https://github.com/tesseract-ocr/tesseract accessed on 05.01.2022) on a sample document image taken from the Marmot dataset [1].

Besides OCR-based approaches, researchers have developed rule-based methods [7–9] that operate on hand-crafted features to tackle the problem of formula detection in documents. Although these methods are effective on a specific type of documents, they are prone to several errors in a diverse nature of the documents. Later, machine learning-based approaches have progressed the benchmark of formula detection systems [10–13]. Very recently, deep neural networks based approaches have been presented in this domain that have remarkably improved the state-of-art of formula detection in scanned document images [13–15]. These approaches have treated the problem of formula identifications as an object detection problem by exploiting modern object detection algorithms [16–19].

In this paper, we take a step forward and formulate the problem of formula identification as an instance segmentation problem. Our system localizes each type of formula (i.e., isolated and embedded) and classifies several instances in a scanned document image. The overall architecture is depicted in Figure 3. In summary, the primary contributions of this paper are as follows:

- We propose an end to end trainable network for formula detection in the scanned document images
- We performed an exhaustive evaluation on ICDAR-2017 and Marmot datasets and achieved state-of-the-art performance on both.

The rest of the paper is organized as follow: Section 2 provides a brief overview of the related approaches for formula localization in scanned document images. Section 3 explains the proposed formula localization framework. This section discusses the deformable convolutions, backbone network, and hybrid task cascade network. Section 4 describes the details of the dataset used for evaluations. Section 5 provides the details of all the experiments and the results of the proposed approach. Furthermore, this section presents the quantitative and qualitative results. Section 6 discusses the conclusion of the proposed work and the possible future directions.



We state and prove next the new work decomposition laws.

THEOREM 5. (WORK DECOMPOSITION LAWS). *Under any dynamic scheduling policy, and for any subset $S \subseteq \mathcal{N}$ of job classes,*

(a)

$$\sum_{j\in S} V_j^S x_j = f(S) + \frac{1}{1-\rho^0(S)} \sum_{i\in S^c} \sum_{j\in S} \lambda_i V_i^S V_j^S x_j^{S_i} + \frac{1-\rho}{1-\rho^0(S)} \sum_{j\in S} V_j^S x_j^0. \tag{30}$$

(b) *Identity* (30) *can be reformulated as*

$$\begin{aligned} E[V^S] = f(S) - \sum_{i\in S} \rho_i(\beta_i - r_i) - \frac{\rho^0(S)}{1-\rho^0(S)} \sum_{i\in S^c} \rho_i r_i \\ + \frac{1}{1-\rho^0(S)} \sum_{i\in S^c} (\lambda_i V_i^S - \rho_i) E^{S_i}[V^S] + \frac{1-\rho(S)}{1-\rho^0(S)} E[V^S | B^S = 0]. \end{aligned} \tag{31}$$

PROOF.

(a) In what follows we use the following notation: if $S, T \subseteq \mathcal{N}$, $\mathbf{z} = (z_i)_{i\in\mathcal{N}}$ is an n-vector, and $\mathbf{A} = (a_{ij})_{i,j\in\mathcal{N}}$ is an $n \times n$ matrix, we shall write

$$\mathbf{z}_S = (z_j)_{j\in S} \quad \text{and} \quad \mathbf{A}_{ST} = (a_{ij})_{i\in S, j\in T}.$$

Let $\mathbf{v}$ denote the n-vector

$$\mathbf{v} = \begin{pmatrix} \mathbf{V}_S^S \\ \mathbf{0} \end{pmatrix},$$

```
We state and prove next the new work decomposition laws.

THEOREM 5. (WORK DECOMPOSITION LAWS). Under any dynamic
scheduling policy, and for
any subset S C N of job classes,
(a)

Si Ip S\.0
(30) Vix; = f(S) + == —s Dd AVIV }x;) + 5 XS) 2 Vix?
JES iE€S* jEs
(b) Identity (30) can be reformulated as
°(S
ELV] = AS) — > piBi-r) - ip Den
(31)
ft Ss S _ p(S ) SI pS —_
ppm & Orr pk] + = aLyBs = 0}

PROOF.

(a) In what follows we use the following notation: if S$, 7 C N, z = (Z;)iey 18 an
n-vector,
and A = (a;);;ey 18 ann X n matrix, we shall write

Zs = (Z,ies and Asr = (dy)iesjer:
Let v denote the n-vector
```

(a) Document image sample.

(b) Retrieved data from OCR.

Figure 2. Figure 2a from the Marmot dataset [1], contains both textual and mathematical information. Figure 2b is the output of an open source Tesseract-OCR [6] employed on Figure 2a. It is evident that the OCR fails to interpret mathematical content.

2. Related work

The problem of mathematical formula detection in documents has been a well studied problem for over two decades [2,7,20,21]. This section discusses the prior approaches that have exploited traditional or deep learning-based methods to tackle the problem of formula detection in documents.

2.1. Traditional Approaches

Initially, similar to the other domains of document image analysis, the task of formula detection has been handled through employing heuristics or rule based methods [21,22]. Fateman et al. [2] published an Optical Character Recognition (OCR) system that can parse mathematical symbols such as $(=, +, cos, sin$, and $x^3)$ in documents. The system works by extracting each character information from documents, and the information is passed to mathematical parsers to output the information as a LaTeXexpression.

Later, researchers have proposed approaches similar to [2] that works on character-based heuristics to extract information from mathematical formulas in documents [23–25]. Inoue et al. [3] classified the mathematical formulas by applying conventional OCR on documents. Since, conventional OCR is unable to parse mathematical formulas, the system treated rest of unrecognized characters as mathematical formulas.

Another fuzzy logic based technique to detect formula region is proposed by Kacem et al. [26]. The authors exploited the features of mathematical symbols to compute the formula boundaries. Jin et al. [7] presented a method to detect isolated and embedded formulas in documents. The authors employ Parzen classifier [27] to detect formulas on the basis of line height and the indentation features.

To detect displayed formulas in PDF documents, the authors in [28] isolates the regular text lines with the formula lines. Similarly, through segeration of regular text lines with the the lines including formulas, Chowdhury et al. [29] incorporate decision trees to predict isolated formulas. Another method is proposed in [8], that tackles the problem of isolated formula detection in documents by exploiting the projection of the features.

Later, researchers have utilized statistical learning specifically machine learning-based algorithms to advance the performance of formula detection systems in documents [7,30]. Liu *et al.* [9] employed the blend of Conditional Random Field (CRF) [31] and Support Vector Machine (SVM) [32] to categorize sparse lines in documents. Then the authors applied heuristics to separate mathematical formulas from other graphical elements in documents.

2.2. Deep Learning Approaches

There has been a noticeable improvement in the field of mathematical formula analysis since the trend of applying vision based algorithms [14]. He et al. [33] introduced the idea of applying Convolutional Neural Networks (CNNs) to extract spatial features in order to detect mathematical formulas in document images. Gao et al. [34] extended the idea of [33] by combining CNN with Recurrent Neural Network (RNN) to leverage from both vision and character level features.

In 2017, a Page Object Detection (POD) competition is arranged at ICDAR [35]. The competition focused on detecting figures, formulas, and tables in document images. NLPR-PAL [35] presented an approach that combines the abilities of Faster R-CNN [16] and connected components to precisely localize the boundaries of figures, formulas, and tables.

Yi *et al.* [36] presented another approach that detects formulas, figures, tables, and text lines in document images. The authors adopt the object detection framework and replaced Non-Maximum Suppression (NMS) with dynamic programming to filter the candidate regions. Ohyama et al. [10] resolved the problem of detecting mathematical expressions through applying U-Net [37] in scientific document images.

Due to the latest improvements of object detection algorithms [17,18,38] in computer vision, there is a growing interest of applying these algorithms in the document image analysis community. Phong *et al.* [11] employed YOLO [17] to identify mathematical expression detection in document images. Furthermore, the authors applied watch, attend, and parse network to extract the content of mathematical expressions.

Along with YOLO, Mali et al. [12] presented the method equipped with SSD [19] to predict mathematical expression in PDF documents. To handle page object detection in PDF document images, Li et al. [39] came up with an hybrid approach that combines traditional and deep learning methods. In a recent work, Younas et al. [13] demonstrated the capabilities of deformable convolutions [40] in Feature Pyramid Networks (FPN) [18] to detect figures and formulas in document images. The presented method depends on works on the transformed images which are achieved by applying traditional computer vision methods. Very recently, Hashmi et al. [15] exploited the combination of composite backbone [41] with cascade Mask R-CNN [38] to improve state-of-the-art results for formula detection in scanned document images.

3. Method

3.1. Deformable Convolution

In the traditional convolutional unit [42], the input feature map is sampled at fixed positions and the output is computed by summing the weighted samples. Hence, the effective receptive field of filters are fixed and decided initially by defining the size of the kernel. Since the formulas present in document images have arbitrary layouts, the conventional convolutional filters do not captures the features well and lack in localizing the precise regions various page objects [13,43,44].

To address the above problem, we propose the novel idea of applying deformable convolution [40] in our ResNeXt backbone and HTC model. Hence, the fixed effective receptive field of the convolutional filters are substituted with the dynamic receptive field due to the presence of a learnable offset that augments the grid. For each location l_0 in the input feature map x, the regular grid is augmented with the help of an offset $\Delta\mathbf{l}_i$. Mathematically, the operation of deformable convolution is explained in [40] as:

$$\mathbf{O}(\mathbf{l}_0) = \sum_{\mathbf{l}_i \in \mathcal{R}} \mathbf{w}(\mathbf{l}_i) \times \mathbf{x}(\mathbf{l}_0 + \mathbf{l}_i + \Delta\mathbf{l}_i) \tag{1}$$

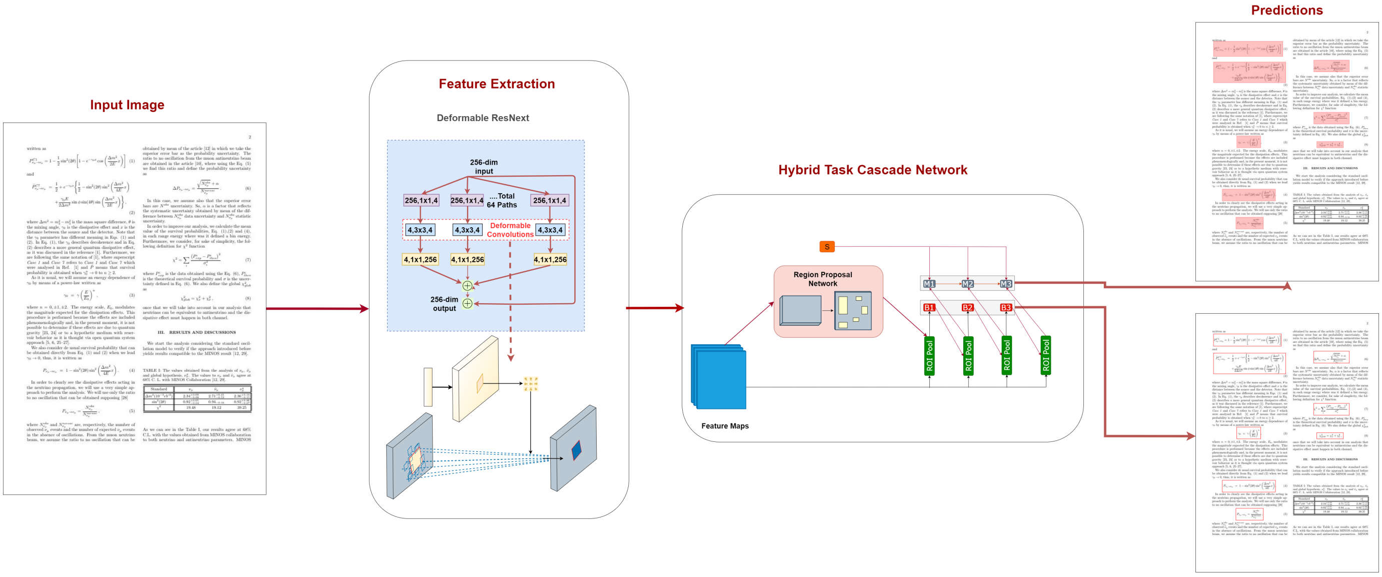


Figure 3. Hybrid Task Cascade Architecture. The feature extraction module depicts that classical convolutions are replaced with deformable convolutions. The block shows the architecture of hybrid task cascade network.

Where l_i enumerates over the regular grid R, w represents the weights, and $O(l_0)$ denotes the output of the feature map on each location l_0. Since the offset $\Delta\mathbf{l}_i$ can update with the backpropagtaion of the gradients, it helps the neurons to adjust the receptive field which facilitates the network to detect formulas with varying scales. Figure 4 exhibits the idea of deformable convolution. We refer our readers to [40,45] for a detailed understanding of deformable convolutions.

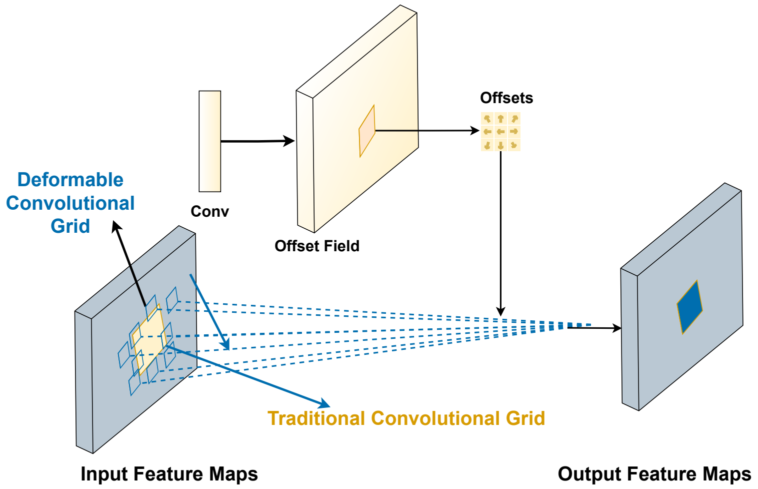


Figure 4. Deformable Convolution.

3.2. HYBRID TASK CASCADE

Cascading has been in use for a very long time in computer vision tasks [38]. It is a generic and robust architecture that helps in achieving better performance. Consequently, this architecture is used to improve the performance of object detection. A naive strategy for implementing the cascading in object detection is to apply iterative bounding box refinement [38]. Although the performance is increased, the improvement is not so much. A hybrid task cascade network introduces a novel way of incorporating the cascade architecture in object detection networks. First, it adopts a fully convolutional branch to provide the spatial context. Secondly, it blends the detection and segmentation tasks

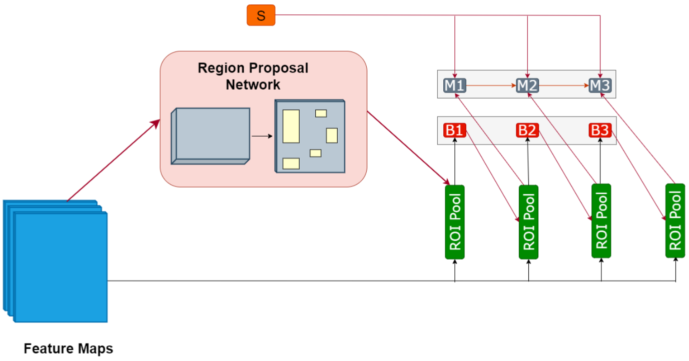


Figure 5. Hybrid Task Cascade Network's architecture. It shows different heads used for predictions. B denotes the bounding box head and M denotes the mask head. Whereas, S represents the

within the cascading framework such that at each cascading stage, we perform both the detection and segmentation. Hence, it is referred to as joint multistage processing. Due to joint multistage processing, object detection and segmentation improve each other. For example, better detection can help to improve the performance of mask prediction and segmentation [46]. The architecture is shown in Figure 3.

The isolated architecture of the hybrid task cascade network is depicted in Figure 5. There exist multiple heads for bounding box prediction and segmentation. Moreover, the input is processed at different scales. The first bounding box head(B1) receives input from the RPN at the first stage, and then the cascade starts, and each subsequent bounding box head receives input from the corresponding ROI align. However, each mask head receives two inputs. The first input comes from the semantic feature maps. The second input comes from the ROI pooling. The mask prediction fuses both of them to generate accurate masks. In summary, the first object proposals come from RPN, which are processed by ROI pooling. The head B1 takes the output of ROI pooling and generates the initial bounding box coordinates. It predicts the confidence of the object proposal as well. M1 generates pixel-wise predictions in terms of masks in the second stage. The rest of the cascaded stages follow the same pattern.

4. Datasets

4.1. ICDAR17

The ICDAR-2017-POD [35] dataset was collected for the competition of graphical page object detection at ICDAR in 2017. The dataset includes information for figures, formulas, and tables in document images. The dataset includes 2417 document images in English language that are retrieved from 1500 scientific papers form CiteSeer. Out of 2417 document images, 1600 images are utilized in training and the remaining 817 samples are used as a test set. The dataset demonstrates a variety of single, double, and multi-column pages having single to several isolated formulas on each image. Recently, Younas et al. [13] pointed out some faulty annotations in the dataset and released the improved version of the dataset. We used the improved version of our dataset in order to have direct comparison with the prior literature.

4.2. Marmot

We evaluate the proposed method on the Marmot dataset [1] as well. This dataset is constructed by extracting PDF documents from CiteSeerX. The dataset contains formula regions for both isolated and embedded formulas in document images. There are 400 document pages retrieved from 194 PDF documents containing 1575 isolated and 7907 embedded formulas. The frequency of isolated and embedded formulas on each sample make this a challenging dataset. For each image, the ground truth is stored in a separate XML file wheres the bounding box of each formula is stored in the hexadecimal numbers.

doi:10.20944/preprints202201.0090.v1

For straight comparisons with previous works [11,47], we used 330 images for training and 70 images for testing.

5. Result and Discussion

We use two datasets ICDAR-2017 POD [35] and Marmot [1] to report the results. This section discusses the qualitative and quantitative results of the proposed approach for both of the datasets mentioned above. Therefore, we will examine both the positive and the negative examples. Moreover, we compare our results with the current state-of-the-art methods.

5.1. ICDAR-17

We use the corrected version of this dataset [13]. Thus, we only compare ourselves with the methods using the same dataset version. We perform the evaluation based on the same protocol as discussed ICDAR-2017 POD [35]. We start the evaluations by computing the test set's true positives, false positives, and false negatives. Then, we translate these raw values into precision, recall and F1-Score as it is illustrated in the previous methods [13,39]. Moreover, we compute mean average precision (mAP) at the same set. Besides these metrics, we also present report IoU at the threshold levels 0.6 and 0.8 as suggested in the competition.

We report the results in Table 1 depicts the results of our approach. It is important to mention that for these results, no pre-processing is used. We achieve a precision of , recall of and f1-score of , and mAP of for the threshold value of 0.6. We achieve better results than the state-of-the-art and the overall achieved accuracy is 96%.

However, the proposed approach achieves much better results in comparison with state-of-the-art methods when the IoU threshold is set to 0.8. We achieve a precision of 0.924, recall of 0.926 and f1-score of 0.925, and mAP 0.96 of for the threshold value of 0.8. We achieve better results than the state-of-the-art. Following the protocol devised by Hashmi et al. [15] we also evaluate the performance on threshold levels of 0.5 and 1.0. Some of the qualitative results of the proposed approach are illustrated in Figures 6 and 7. Whereas, Figure 8 depicts f1-score for all of the above mentioned threshold.

Comparison with State-of-the-art Methods

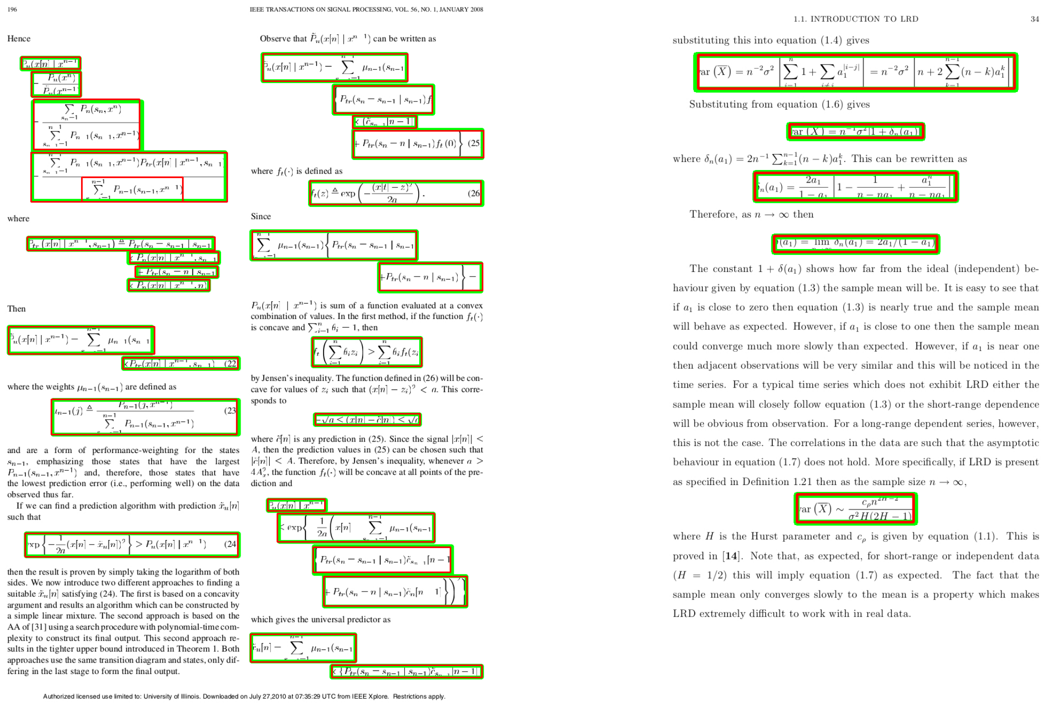
As reported in Table 1, the proposed approach outperforms the state-of-the-art methods [15]. We obtain an f1-score of 0.975, resulting in a relative error improvement of 8.5% for IoU of 0.6. However, when it comes to more strict threshold that is 0.8, the relative error rate is much decreased. We acheive an f1 score of 0.96 with that results in relative error reduction of 13%.

5.2. Marmot

As per the convention we followed for ICDAR-2017 POD dataset, we follow the same protocol as the previous approaches for evaluating the Marmot dataset. It enables us to compare our results directly with the other approaches. It is essential to mention that we detect the isolated and the embedded formulas separately. We outline the quantitative results of our approach in the Table 2. We compute the accuracy of both complete and partial detections. The proposed approach achieves the correct detection accuracy of 93.5%. For embedded formulas, the proposed approach obtains the correct detection accuracy of 82.1%. We evaluate the performance over a range of IoU threshold, starting from 0.5 to 1.0 as highlighted in Figure 9. As far as the qualitative performance is concerned, the Figures 10, 11, and 12 depicts the performance of the proposed approach.

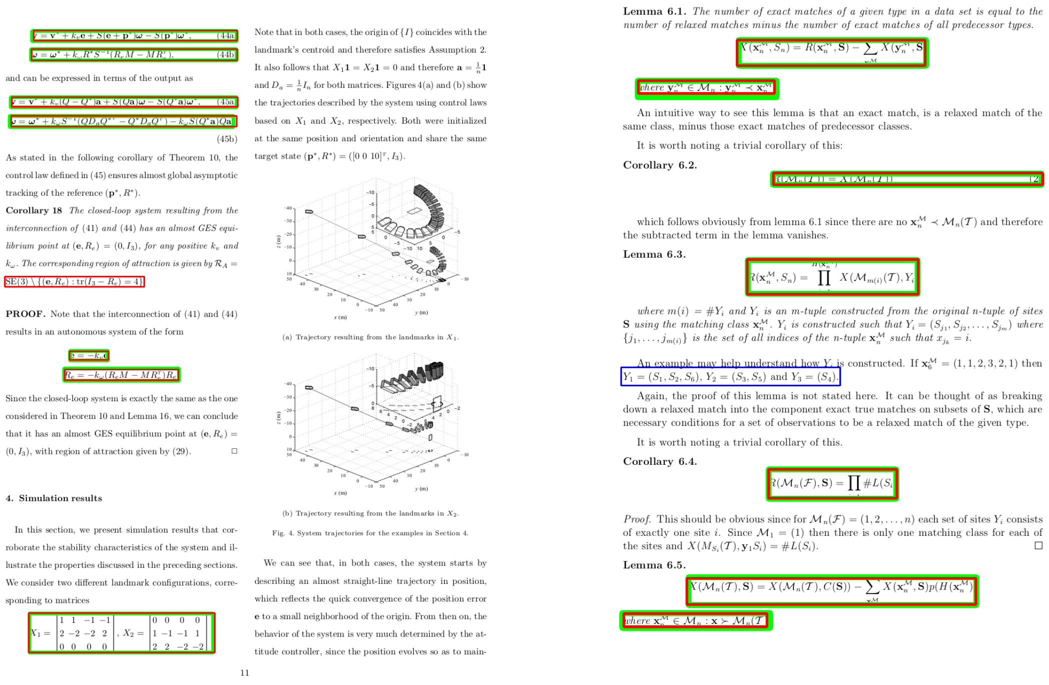
Comparison with State-of-the-Art Methods

The comparison of our results with earlier approaches on the Marmot dataset is depicted in Table 2. We outperform the current state-of-the-art with a good margin. For embedded formula, we achieve an accuracy of 82.1%, thereby reducing the relative



(a) True positives on a two-column document image. **(b)** True positives on a single column document image.

Figure 6. Isolated Formula detection results on the ICDAR-2017-POD dataset. The green colour depicts the ground truth, while red denotes predictions. Figures 6a and 6b exhibit instances of true positives.



(a) True positives and a false positive. **(b)** False positives and a false negative.

Figure 7. Instances of faulty predictions on the ICDAR-2017-POD dataset. The green colour depicts the ground truth, while red denotes the predicted bounding boxes. Figure 7a contains a single false positive, while Figure 7b illustrates a single false negative.

doi:10.20944/preprints202201.0090.v1

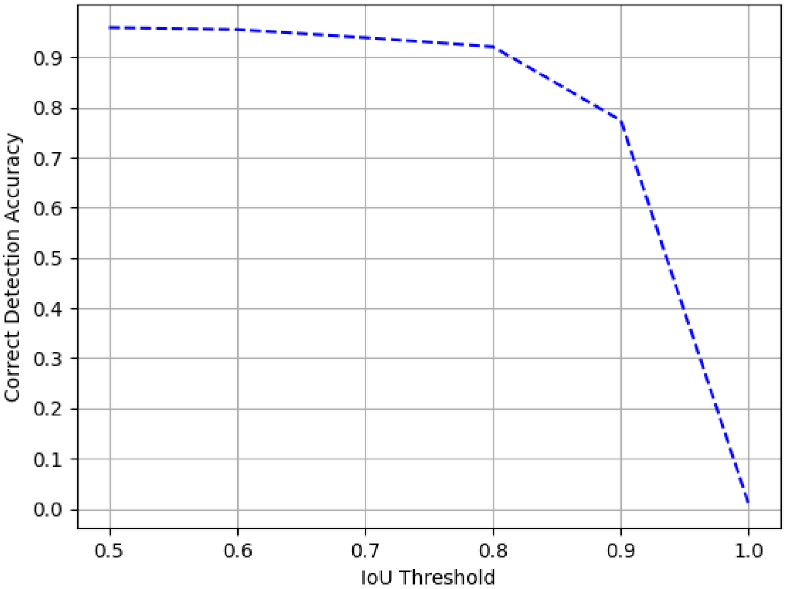
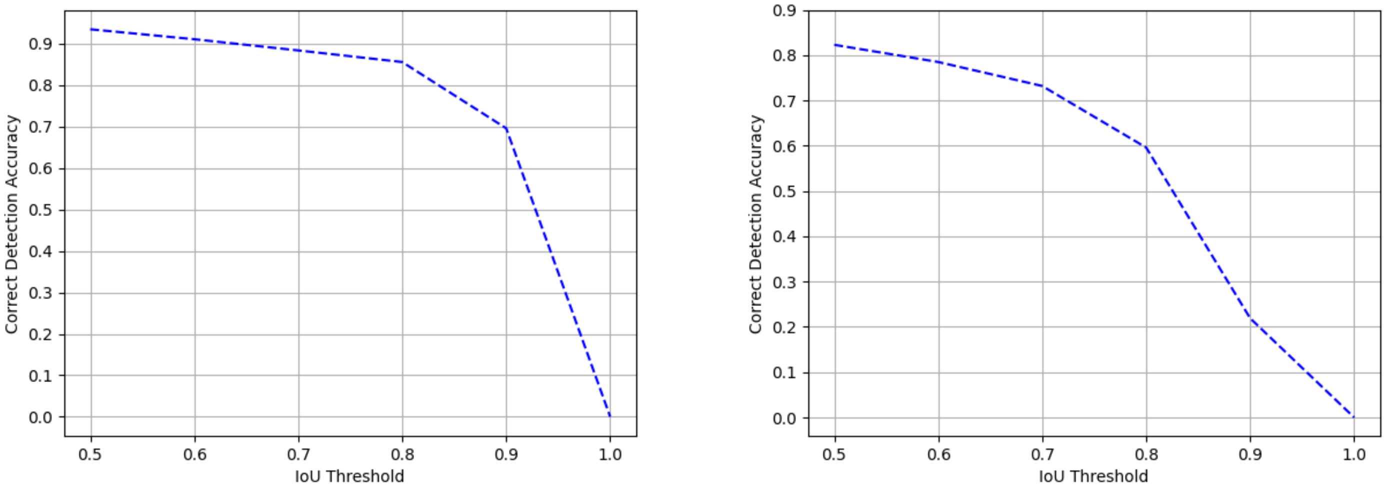


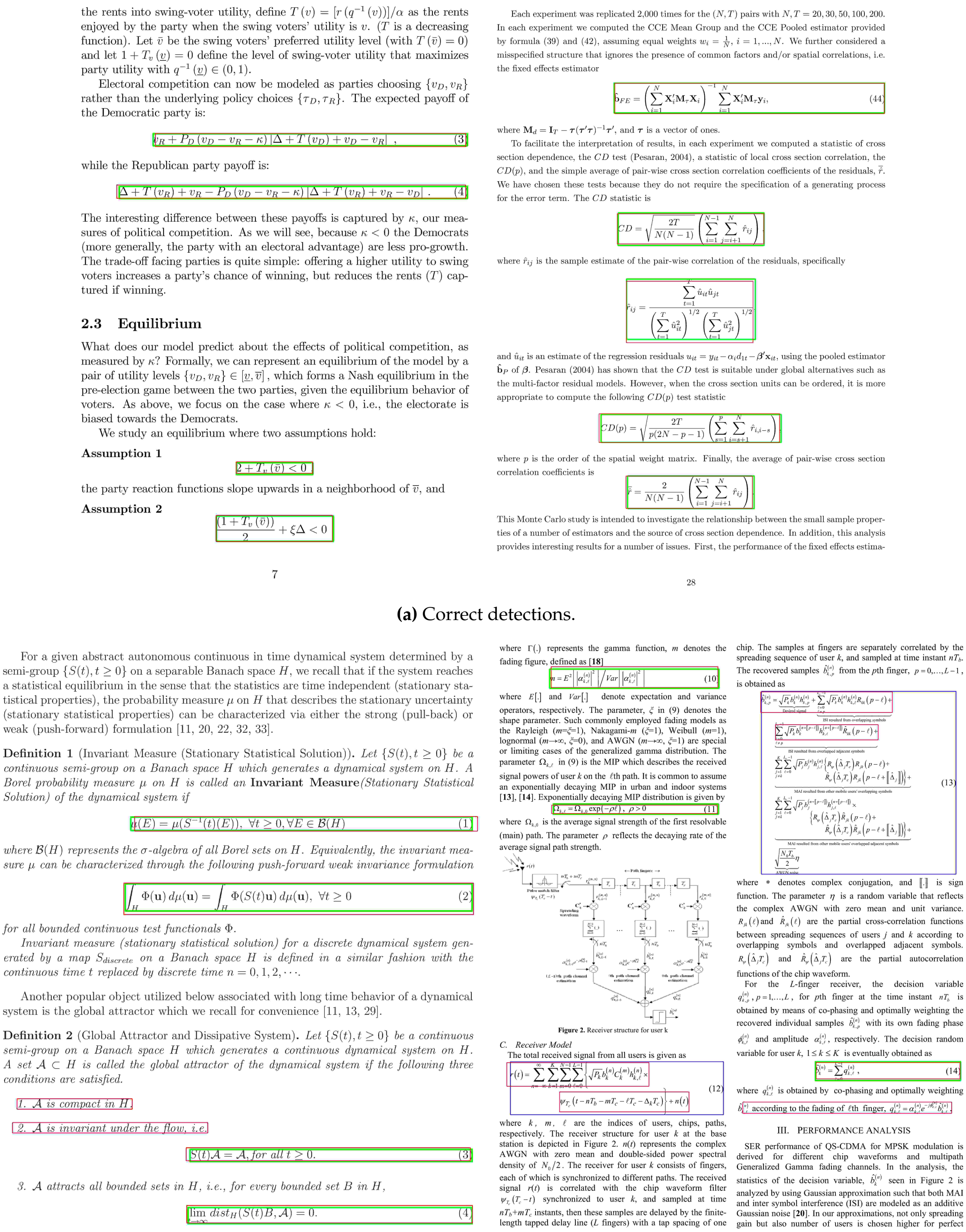
Figure 8. F1-score of our method on ICDAR 17 POD Dataset on IoU ranging from 0.5 - 1.0.



(a) Isolated formula detection.

(b) Embedded formula detection.

Figure 9. Illustrating the decline of detection accuracy on increasing IoU threshold from 0.5 to 1.0. Figure 9a highlights accuracy on isolated formulas, whereas Figure 9b shows an accuracy on embedded formulas.



the rents into swing-voter utility, define $T(v) = [r(q^{-1}(v))]/\alpha$ as the rents enjoyed by the party when the swing voters' utility is v. (T is a decreasing function). Let $\bar{v}$ be the swing voters' preferred utility level (with $T(\bar{v}) = 0$) and let $1 + T_v(\underline{v}) = 0$ define the level of swing-voter utility that maximizes party utility with $q^{-1}(\underline{v}) \in (0, 1)$.

Electoral competition can now be modeled as parties choosing $\{v_D, v_R\}$ rather than the underlying policy choices $\{\tau_D, \tau_R\}$. The expected payoff of the Democratic party is:

$$v_R + P_D(v_D - v_R - \kappa)[\Delta + T(v_D) + v_D - v_R] \ , \tag{3}$$

while the Republican party payoff is:

$$[\Delta + T(v_R) + v_R - P_D(v_D - v_R - \kappa)[\Delta + T(v_R) + v_R - v_D] \ . \tag{4}$$

The interesting difference between these payoffs is captured by κ, our measures of political competition. As we will see, because $\kappa < 0$ the Democrats (more generally, the party with an electoral advantage) are less pro-growth. The trade-off facing parties is quite simple: offering a higher utility to swing voters increases a party's chance of winning, but reduces the rents (T) captured if winning.

2.3 Equilibrium

What does our model predict about the effects of political competition, as measured by κ? Formally, we can represent an equilibrium of the model by a pair of utility levels $\{v_D, v_R\} \in [\underline{v}, \bar{v}]$, which forms a Nash equilibrium in the pre-election game between the two parties, given the equilibrium behavior of voters. As above, we focus on the case where $\kappa < 0$, i.e., the electorate is biased towards the Democrats.

We study an equilibrium where two assumptions hold:

Assumption 1

$$2 + T_v(\bar{v}) < 0$$

the party reaction functions slope upwards in a neighborhood of $\bar{v}$, and

Assumption 2

$$\frac{(1 + T_v(\bar{v}))}{2} + \xi\Delta < 0$$

7

Each experiment was replicated 2,000 times for the (N, T) pairs with $N, T = 20, 30, 50, 100, 200$. In each experiment we computed the CCE Mean Group and the CCE Pooled estimator provided by formula (39) and (42), assuming equal weights $w_i = \frac{1}{N}$, $i = 1, ..., N$. We further considered a misspecified structure that ignores the presence of common factors and/or spatial correlations, i.e. the fixed effects estimator

$$\hat{\mathbf{b}}_{FE} = \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{M}_\tau \mathbf{X}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{M}_\tau \mathbf{y}_i, \tag{44}$$

where $\mathbf{M}_d = \mathbf{I}_T - \boldsymbol{\tau}(\boldsymbol{\tau}'\boldsymbol{\tau})^{-1}\boldsymbol{\tau}'$, and $\boldsymbol{\tau}$ is a vector of ones.

To facilitate the interpretation of results, in each experiment we computed a statistic of cross section dependence, the CD test (Pesaran, 2004), a statistic of local cross section correlation, the $CD(p)$, and the simple average of pair-wise cross section correlation coefficients of the residuals, $\bar{\hat{r}}$. We have chosen these tests because they do not require the specification of a generating process for the error term. The CD statistic is

$$CD = \sqrt{\frac{2T}{N(N-1)}} \left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{r}_{ij}\right)$$

where $\hat{r}_{ij}$ is the sample estimate of the pair-wise correlation of the residuals, specifically

$$\hat{r}_{ij} = \frac{\sum_{t=1}^{T} \hat{u}_{it}\hat{u}_{jt}}{\left(\sum_{t=1}^{T} \hat{u}_{it}^2\right)^{1/2} \left(\sum_{t=1}^{T} \hat{u}_{jt}^2\right)^{1/2}}$$

and $\hat{u}_{it}$ is an estimate of the regression residuals $u_{it} = y_{it} - \alpha_i d_{1t} - \boldsymbol{\beta}'\mathbf{x}_{it}$, using the pooled estimator $\hat{\mathbf{b}}_P$ of $\boldsymbol{\beta}$. Pesaran (2004) has shown that the CD test is suitable under global alternatives such as the multi-factor residual models. However, when the cross section units can be ordered, it is more appropriate to compute the following $CD(p)$ test statistic

$$CD(p) = \sqrt{\frac{2T}{p(2N-p-1)}} \left(\sum_{s=1}^{p} \sum_{i=s+1}^{N} \hat{r}_{i,i-s}\right)$$

where p is the order of the spatial weight matrix. Finally, the average of pair-wise cross section correlation coefficients is

$$\bar{\hat{r}} = \frac{2}{N(N-1)} \left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{r}_{ij}\right)$$

This Monte Carlo study is intended to investigate the relationship between the small sample properties of a number of estimators and the source of cross section dependence. In addition, this analysis provides interesting results for a number of issues. First, the performance of the fixed effects estima-

28

(a) Correct detections.

For a given abstract autonomous continuous in time dynamical system determined by a semi-group $\{S(t), t \geq 0\}$ on a separable Banach space H, we recall that if the system reaches a statistical equilibrium in the sense that the statistics are time independent (stationary statistical properties), the probability measure μ on H that describes the stationary uncertainty (stationary statistical properties) can be characterized via either the strong (pull-back) or weak (push-forward) formulation [11, 20, 22, 32, 33].

Definition 1 (Invariant Measure (Stationary Statistical Solution)). *Let $\{S(t), t \geq 0\}$ be a continuous semi-group on a Banach space H which generates a dynamical system on H. A Borel probability measure μ on H is called an* **Invariant Measure***(Stationary Statistical Solution) of the dynamical system if*

$$\mu(E) = \mu(S^{-1}(t)(E)), \ \forall t \geq 0, \forall E \in \mathcal{B}(H) \tag{1}$$

where $\mathcal{B}(H)$ represents the σ-algebra of all Borel sets on H. Equivalently, the invariant measure μ can be characterized through the following push-forward weak invariance formulation

$$\int_H \Phi(\mathbf{u})\, d\mu(\mathbf{u}) = \int_H \Phi(S(t)\mathbf{u})\, d\mu(\mathbf{u}), \ \forall t \geq 0 \tag{2}$$

for all bounded continuous test functionals Φ.

Invariant measure (stationary statistical solution) for a discrete dynamical system generated by a map $S_{discrete}$ on a Banach space H is defined in a similar fashion with the continuous time t replaced by discrete time $n = 0, 1, 2, \cdots$.

Another popular object utilized below associated with long time behavior of a dynamical system is the global attractor which we recall for convenience [11, 13, 29].

Definition 2 (Global Attractor and Dissipative System). *Let $\{S(t), t \geq 0\}$ be a continuous semi-group on a Banach space H which generates a continuous dynamical system on H. A set $\mathcal{A} \subset H$ is called the global attractor of the dynamical system if the following three conditions are satisfied.*

1. $\mathcal{A}$ is compact in H.

2. $\mathcal{A}$ is invariant under the flow, i.e.

$$S(t)\mathcal{A} = \mathcal{A}, \textit{for all } t \geq 0. \tag{3}$$

3. $\mathcal{A}$ attracts all bounded sets in H, i.e., for every bounded set B in H,

$$\lim_{t\to\infty} dist_H(S(t)B, \mathcal{A}) = 0. \tag{4}$$

where $\Gamma(.)$ represents the gamma function, m denotes the fading figure, defined as [**18**]

$$m = E^2\left[\left|\alpha_{k,\ell}^{(n)}\right|^2\right] \Big/ Var\left[\left|\alpha_{k,\ell}^{(n)}\right|^2\right] \tag{10}$$

where $E[.]$ and $Var[.]$ denote expectation and variance operators, respectively. The parameter, ξ in (9) denotes the shape parameter. Such commonly employed fading models as the Rayleigh (m=ξ=1), Nakagami-m (ξ=1), Weibull (m=1), lognormal (m→∞, ξ=0), and AWGN (m→∞, ξ=1) are special or limiting cases of the generalized gamma distribution. The parameter $\Omega_{k,\ell}$ in (9) is the MIP which describes the received signal powers of user k on the ℓth path. It is common to assume an exponentially decaying MIP in urban and indoor systems [**13**], [**14**]. Exponentially decaying MIP distribution is given by

$$\Omega_{k,\ell} = \Omega_{k,0}\exp(-\rho\ell), \ \rho > 0 \tag{11}$$

where $\Omega_{k,0}$ is the average signal strength of the first resolvable (main) path. The parameter ρ reflects the decaying rate of the average signal path strength.

Figure 2. Receiver structure for user k

C. Receiver Model

The total received signal from all users is given as

$$r(t) = \sum_{n=-\infty}^{\infty}\sum_{k=1}^{K}\sum_{m=0}^{N-1}\sum_{\ell=0}^{L-1}\Big\{\sqrt{P_k}\, b_k^{(n)} C_k^{(m)} h_{k,\ell}^{(n)} \times \psi_{T_c}\left(t - nT_b - mT_c - \ell T_c - \Delta_k T_c\right)\Big\} + n(t) \tag{12}$$

where k, m, ℓ are the indices of users, chips, paths, respectively. The receiver structure for user k at the base station is depicted in Figure 2. $n(t)$ represents the complex AWGN with zero mean and double-sided power spectral density of $N_0/2$. The receiver for user k consists of fingers, each of which is synchronized to different paths. The received signal $r(t)$ is correlated with the chip waveform filter $\psi_{T_c}(T_c - t)$ synchronized to user k, and sampled at time nT_b+mT_c instants, then these samples are delayed by the finite-length tapped delay line (L fingers) with a tap spacing of one chip. The samples at fingers are separately correlated by the spreading sequence of user k, and sampled at time instant nT_b. The recovered samples $\hat{b}_{k,p}^{(n)}$ from the pth finger, $p = 0,\ldots,L-1$, is obtained as

$$\begin{aligned}
\hat{b}_{k,p}^{(n)} = {} & \underbrace{\sqrt{P_k}\, b_k^{(n)} h_{k,p}^{(n)}}_{\text{Desired signal}} + \underbrace{\sum_{\substack{\ell=0\\ \ell\neq p}}^{L_c-1}\sqrt{P_k}\, b_k^{(n)} h_{k,\ell}^{(n)} R_{kk}(p-\ell)}_{\text{ISI resulted from overlapping symbols}} + \\
& \underbrace{\sum_{\substack{\ell=0\\ \ell\neq p}}^{L_c-1}\sqrt{P_k}\, b_k^{(n+[\![p-\ell]\!])} h_{k,\ell}^{(n+[\![p-\ell]\!])}\hat{R}_{kk}(p-\ell)}_{\text{ISI resulted from overlapped adjacent symbols}} + \\
& \underbrace{\sum_{\substack{j=1\\ j\neq k}}^{K}\sum_{\ell=0}^{L_c-1}\sqrt{P_j}\, b_j^{(n)} h_{j,\ell}^{(n)}\Big\{R_\psi(\hat{\Delta}_j T_c) R_{jk}(p-\ell) + \hat{R}_\psi(\hat{\Delta}_j T_c) R_{jk}\left(p-\ell+[\![\hat{\Delta}_j]\!]\right)\Big\}}_{\text{MAI resulted from other mobile users' overlapping symbols}} + \\
& \underbrace{\sum_{\substack{j=1\\ j\neq k}}^{K}\sum_{\ell=0}^{L_c-1}\sqrt{P_j}\, b_j^{(n+[\![p-\ell]\!])} h_{j,\ell}^{(n+[\![p-\ell]\!])} \times \Big\{R_\psi(\hat{\Delta}_j T_c)\hat{R}_{jk}(p-\ell) + \hat{R}_\psi(\hat{\Delta}_j T_c)\hat{R}_{jk}\left(p-\ell+[\![\hat{\Delta}_j]\!]\right)\Big\}}_{\text{MAI resulted from other mobile users' overlapped adjacent symbols}} + \\
& \underbrace{\sqrt{\frac{N_0 T_b}{2}}\eta}_{\text{AWGN noise}}
\end{aligned} \tag{13}$$

where $*$ denotes complex conjugation, and $[\![.]\!]$ is sign function. The parameter η is a random variable that reflects the complex AWGN with zero mean and unit variance. $R_{jk}(\ell)$ and $\hat{R}_{jk}(\ell)$ are the partial cross-correlation functions between spreading sequences of users j and k according to overlapping symbols and overlapped adjacent symbols. $R_\psi(\hat{\Delta}_j T_c)$ and $\hat{R}_\psi(\hat{\Delta}_j T_c)$ are the partial autocorrelation functions of the chip waveform.

For the L-finger receiver, the decision variable $q_{k,p}^{(n)}$, $p = 1,\ldots,L$, for pth finger at the time instant nT_b is obtained by means of co-phasing and optimally weighting the recovered individual samples $\hat{b}_{k,p}^{(n)}$ with its own fading phase $\phi_{k,\ell}^{(n)}$ and amplitude $\alpha_{k,\ell}^{(n)}$, respectively. The decision random variable for user k, $1 \leq k \leq K$ is eventually obtained as

$$\hat{b}_k^{(n)} = \sum_{\ell=0}^{L-1} q_{k,\ell}^{(n)}, \tag{14}$$

where $q_{k,\ell}^{(n)}$ is obtained by co-phasing and optimally weighting $\hat{b}_{k,\ell}^{(n)}$ according to the fading of ℓth finger, $q_{k,\ell}^{(n)} = \alpha_{k,\ell}^{(n)} e^{-j\phi_{k,\ell}^{(n)}}\hat{b}_{k,\ell}^{(n)}$.

III. PERFORMANCE ANALYSIS

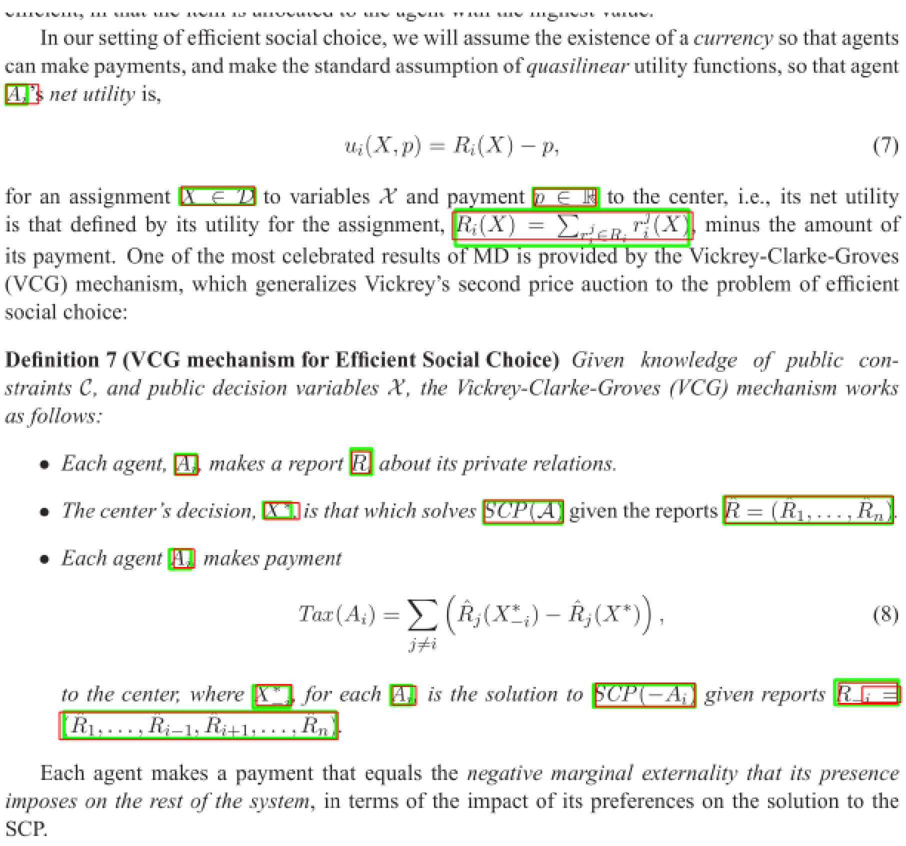
SER performance of QS-CDMA for MPSK modulation is derived for different chip waveforms and multipath Generalized Gamma fading channels. In the analysis, the statistics of the decision variable, $\hat{b}_k^{(n)}$ seen in Figure 2 is analyzed by using Gaussian approximation such that both MAI and inter symbol interference (ISI) are modeled as an additive Gaussian noise [**20**]. In our approximations, not only spreading gain but also number of users is chosen higher for perfect

(b) Partial and missed detections.

Figure 10. Qualitative results on the Marmot dataset. The greens represents the correct detections, red and blue highlight incorrect and missed detection, respectively.

Table 1. Comparison with existing state-of-the-art methods on the ICDAR-2017 POD dataset.

ICDAR-2017 POD								
Method	**IoU = 0.6**				**IoU = 0.8**			
	Precision	**Recall**	**F1-Score**	**AP**	**Precision**	**Recall**	**F1-Score**	**AP**
NLPR-PAL [35]	0.901	0.929	0.915	0.839	0.888	0.916	0.902	0.816
Li et al. [39]	0.935	0.331	0.489	0.312	0.877	0.310	0.459	0.274
Fi-Fo Detector Non Deformable [13]	0.910	0.927	0.918	0.953	0.860	0.877	0.868	0.928
Fi-Fo Detector Deformable [13]	0.957	0.952	0.954	0.949	0.913	0.908	0.910	0.898
Hashmi et al. [15]	0.954	**0.952**	0.953	0.970	0.918	0.916	0.917	0.954
Our Method	**0.961**	**0.954**	**0.957**	**0.975**	**0.924**	**0.926**	**0.925**	**0.960**



efficient, in that the item is allocated to the agent with the highest value.

In our setting of efficient social choice, we will assume the existence of a *currency* so that agents can make payments, and make the standard assumption of *quasilinear* utility functions, so that agent A_i's *net utility* is,

$$u_i(X, p) = R_i(X) - p, \qquad (7)$$

for an assignment $X \in \mathcal{D}$ to variables $\mathcal{X}$ and payment $p \in \mathbb{R}$ to the center, i.e., its net utility is that defined by its utility for the assignment, $R_i(X) = \sum_{r_i^j \in R_i} r_i^j(X)$, minus the amount of its payment. One of the most celebrated results of MD is provided by the Vickrey-Clarke-Groves (VCG) mechanism, which generalizes Vickrey's second price auction to the problem of efficient social choice:

Definition 7 (VCG mechanism for Efficient Social Choice) *Given knowledge of public constraints $\mathcal{C}$, and public decision variables $\mathcal{X}$, the Vickrey-Clarke-Groves (VCG) mechanism works as follows:*

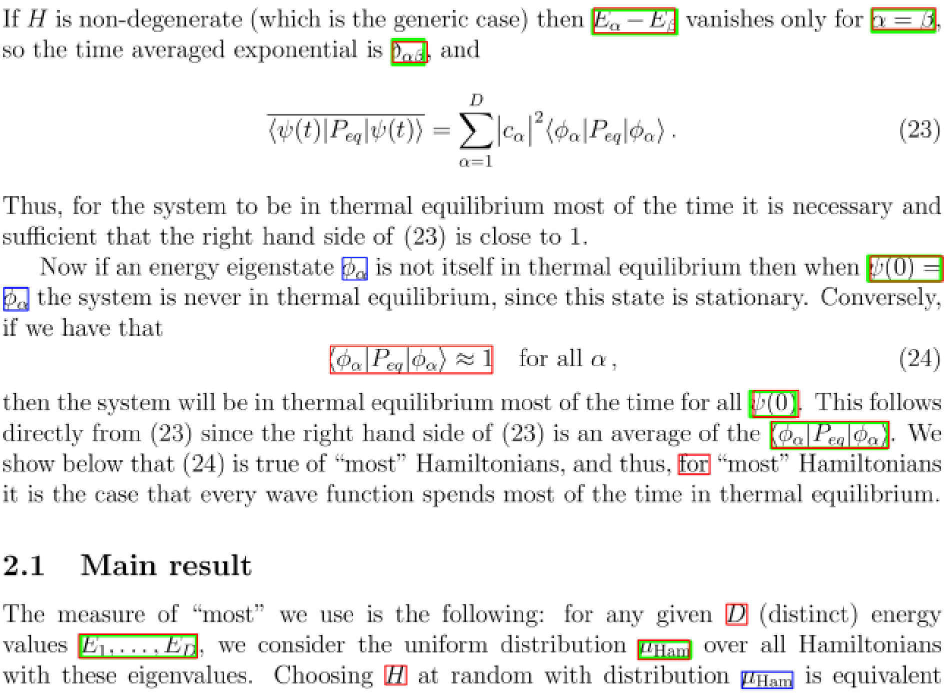
- *Each agent, A_i, makes a report $\hat{R}_i$ about its private relations.*
- *The center's decision, X^*, is that which solves $SCP(\mathcal{A})$* given the reports $\hat{R} = (\hat{R}_1, \ldots, \hat{R}_n)$.
- *Each agent A_i makes payment*

$$Tax(A_i) = \sum_{j \neq i} \left(\hat{R}_j(X^*_{-i}) - \hat{R}_j(X^*) \right), \qquad (8)$$

to the center, where X^*_{-i}, for each A_i, is the solution to $SCP(-A_i)$ given reports $\hat{R}_{-i} = (\hat{R}_1, \ldots, \hat{R}_{i-1}, \hat{R}_{i+1}, \ldots, \hat{R}_n)$.

Each agent makes a payment that equals the *negative marginal externality that its presence imposes on the rest of the system*, in terms of the impact of its preferences on the solution to the SCP.

Figure 11. Visualization of correct detections of embedded formulas. For better understanding, a fragment of document images is taken from the marmot dataset. The green shows ground truth, whereas red displays the predictions.



If H is non-degenerate (which is the generic case) then $E_\alpha - E_\beta$ vanishes only for $\alpha = \beta$, so the time averaged exponential is $\delta_{\alpha\beta}$, and

$$\overline{\langle\psi(t)|P_{eq}|\psi(t)\rangle} = \sum_{\alpha=1}^{D} |c_\alpha|^2 \langle\phi_\alpha|P_{eq}|\phi_\alpha\rangle\,. \tag{23}$$

Thus, for the system to be in thermal equilibrium most of the time it is necessary and sufficient that the right hand side of (23) is close to 1.

Now if an energy eigenstate ϕ_α is not itself in thermal equilibrium then when $\psi(0) = \phi_\alpha$ the system is never in thermal equilibrium, since this state is stationary. Conversely, if we have that

$$\langle\phi_\alpha|P_{eq}|\phi_\alpha\rangle \approx 1 \quad \text{for all } \alpha\,, \tag{24}$$

then the system will be in thermal equilibrium most of the time for all $\psi(0)$. This follows directly from (23) since the right hand side of (23) is an average of the $\langle\phi_\alpha|P_{eq}|\phi_\alpha\rangle$. We show below that (24) is true of "most" Hamiltonians, and thus, for "most" Hamiltonians it is the case that every wave function spends most of the time in thermal equilibrium.

2.1 Main result

The measure of "most" we use is the following: for any given D (distinct) energy values $E_1, \ldots, E_D$, we consider the uniform distribution μ_{Ham} over all Hamiltonians with these eigenvalues. Choosing H at random with distribution μ_{Ham} is equivalent

Figure 12. Partial and missed detections in case of embedded formulas on the Marmot dataset. Green highlights correct predictions, partial and missed detections are depicted with red and blue, respectively.

error by 4.3%. Furthermore, the rate of partial detection is improved from 4.86% [15] to 5.28% using the proposed method. This, in turn, increases the overall accuracy from 97.86% [15] to 98.78%. Thus we achieve an overall error reduction of 43% for isolated formula. In the case of embedded formulas, we achieve an accuracy of 82.1%. However, the partial detection rate improves from 6.77% [15] to 8.11%. As a result, we achieve an overall accuracy of 90.21%. This resulted in the overall reduction of the error rate of 17.9%. These results further highlight the supremacy of the proposed approach.

Table 2. Quantitative analysis between our method and previous state-of-the-art approaches on the Marmot dataset.

Method	Formula	Correct (%)	Partial (%)	Total
Chu et al. [48]	Isolated	26.87	44.87	71.76
	Embedded	1.74	28.87	30.61
Phong et al. [47]	Isolated	50.37	39.14	91.18
	Embedded	22.9	58.45	81.35
Phong et al. [11]	Isolated	93	-	-
	Embedded	73	-	-
Hashmi et al. [15]	Isolated	93	4.86	97.86
	Embedded	81.3	6.77	88.07
Our method	Isolated	**93.5**	**5.28**	**98.78**
	Embedded	**82.1**	**8.11**	**90.21**

6. Conclusion and Future Work

We proposed an end-to-end trainable network to localize embedded and isolated formulas in the scanned document images. We replace classical convolutions with deformable convolutions. Furthermore, we used the backbone ResNext101. Both of these modifications result in the superior performance of the network. We evaluate the proposed method on ICDAR2017-POD and Marmot datasets. We report both the qualitative and quantitative results for the proposed approach. We observe a significant improvement in the relative error rate both for the embedded and the isolated formulas, thereby validating the superiority of the proposed approach. Furthermore, we observe that the proposed approach performs significantly better with a higher threshold than the current state of the art.

One future direction to extend the work is the introduction of a rather powerful backbone. We expect a performance gain if the backbone is more profound and can provide refined proposals. Furthermore, other graphical objects such as charts, figures, and tables [49] can benefit from the proposed work.

Author Contributions: writing—original draft preparation, M.Z.A., K.A.H.; writing—review and editing, M.Z.A., K.A.H., M.L.; supervision and project administration, A.P., D.S. All authors have read and agreed to the submitted version of the manuscript.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fang, J.; Tao, X.; Tang, Z.; Qiu, R.; Liu, Y. Dataset, ground-truth and performance metrics for table detection evaluation. 2012 10th IAPR International Workshop on Document Analysis Systems. IEEE, March 27–29, 2012; Queensland, Australia, pp. 445–449.
2. Fateman, R.J.; Tokuyasu, T.; Berman, B.P.; Mitchell, N. Optical character recognition and parsing of typeset mathematics1. *Journal of Visual Communication and Image Representation* **1996**, *7*, 2–15.
3. Inoue, K.; Miyazaki, R.; Suzuki, M. Optical recognition of printed mathematical documents. Proc. Third Asian Technology Conf. Math, 1998, Vol. 3, pp. 280–289.
4. Kieninger, T.; Dengel, A. The t-recs table recognition and analysis system. International Workshop on Document Analysis Systems. Springer, 1998, pp. 255–270.
5. Hashmi, K.A.; Ponnappa, R.B.; Bukhari, S.S.; Jenckel, M.; Dengel, A. Feedback Learning: Automating the Process of Correcting and Completing the Extracted Information. 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). IEEE, September 20–25, 2019, Sydney, Australia, Vol. 5, pp. 116–121.
6. Smith, R. An overview of the Tesseract OCR engine. Ninth international conference on document analysis and recognition (ICDAR 2007). IEEE, September 23–26, 2007, Curitiba, Brazil, Vol. 2, pp. 629–633.
7. Jin, J.; Han, X.; Wang, Q. Mathematical Formulas Extraction. Icdar. Citeseer, 2003, pp. 1138–1141.
8. Chang, T.Y.; Takiguchi, Y.; Okada, M. Physical structure segmentation with projection profile for mathematic formulae and graphics in academic paper images. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). IEEE, September 23–26, 2007; State of Paraná, Brazil, Vol. 2, pp. 1193–1197.
9. Liu, Y.; Bai, K.; Gao, L. An efficient pre-processing method to identify logical components from pdf documents. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2011, pp. 500–511.
10. Ohyama, W.; Suzuki, M.; Uchida, S. Detecting mathematical expressions in scientific document images using a u-net trained on a diverse dataset. *IEEE Access* **2019**, *7*, 144030–144042.
11. Phong, B.H.; Dat, L.T.; Yen, N.T.; Hoang, T.M.; Le, T.L. A deep learning based system for mathematical expression detection and recognition in document images. 2020 12th International Conference on Knowledge and Systems Engineering (KSE). IEEE, November 12–14, 2020; Can Tho City, Vietnam, pp. 85–90.
12. Mali, P.; Kukkadapu, P.; Mahdavi, M.; Zanibbi, R. ScanSSD: Scanning Single Shot Detector for Mathematical Formulas in PDF Document Images. *arXiv preprint arXiv:2003.08005* **2020**.
13. Younas, J.; Siddiqui, S.A.; Munir, M.; Malik, M.I.; Shafait, F.; Lukowicz, P.; Ahmed, S. Fi-Fo Detector: Figure and Formula Detection Using Deformable Networks. *Applied Sciences* **2020**, *10*, 6460.
14. Bhatt, J.; Hashmi, K.; Afzal, M.; Stricker, D. A Survey of Graphical Page Object Detection with Deep Neural Networks. *arXiv* **2021**.
15. Hashmi, K.A.; Pagani, A.; Liwicki, M.; Stricker, D.; Afzal, M.Z. Cascade network with deformable composite backbone for formula detection in scanned document images. *Applied Sciences* **2021**, *11*, 7610.

16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* **2015**.
17. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* **2018**.
18. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection, 2017, [arXiv:cs.CV/1612.03144].
19. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. European conference on computer vision. Springer, October 8–16, 2016, Berlin/Heidelberg, Germany, pp. 21–37.
20. Lin, X.; Gao, L.; Tang, Z.; Baker, J.; Sorge, V. Mathematical formula identification and performance evaluation in PDF documents. *International Journal on Document Analysis and Recognition (IJDAR)* **2014**, *17*, 239–255.
21. Chan, K.F.; Yeung, D.Y. Mathematical expression recognition: a survey. *International Journal on Document Analysis and Recognition* **2000**, *3*, 3–15.
22. Zanibbi, R.; Blostein, D. Recognition and retrieval of mathematical expressions. *International Journal on Document Analysis and Recognition (IJDAR)* **2012**, *15*, 331–357.
23. Lee, H.J.; Wang, J.S. Design of a mathematical expression understanding system. *Pattern Recognition Letters* **1997**, *18*, 289–298.
24. Toumit, J.Y.; Garcia-Salicetti, S.; Emptoz, H. A hierarchical and recursive model of mathematical expressions for automatic reading of mathematical documents. Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318). IEEE, September 20–22, 1999; Bangalore, India, pp. 119–122.
25. Garain, U.; Chaudhuri, B. A syntactic approach for processing mathematical expressions in printed documents. Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. IEEE, September 3–7, 2000; Barcelona, Spain, Vol. 4, pp. 523–526.
26. Kacem, A.; Belaïd, A.; Ahmed, M.B. Automatic extraction of printed mathematical formulas using fuzzy logic and propagation of context. *International Journal on Document Analysis and Recognition* **2001**, *4*, 97–108.
27. Fukunaga, K.; Hayes, R.R. The reduced Parzen classifier. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1989**, *11*, 423–425.
28. Baker, J.B.; Sexton, A.P.; Sorge, V. Towards Reverse Engineering of PDF Documents. *DML 2011 Towards a Digital Mathematics Library* **2011**, *4*, 65–75.
29. Chowdhury, S.; Mandal, S.; Das, A.K.; Chanda, B. Automated segmentation of math-zones from document images. Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. Citeseer, August 3–6, 2003; Edinburgh, Scotland, UK, pp. 755–759.
30. Drake, D.M.; Baird, H.S. Distinguishing mathematics notation from English text using computational geometry. Eighth international conference on document analysis and recognition (ICDAR'05). IEEE, August 29–September 1, 2005; Seoul, Korea, pp. 1270–1274.
31. Tseng, H.; Chang, P.C.; Andrew, G.; Jurafsky, D.; Manning, C.D. A conditional random field word segmenter for sighan bakeoff 2005. Proceedings of the fourth SIGHAN workshop on Chinese language Processing, 2005.
32. Schölkopf, B.; Williamson, R.C.; Smola, A.J.; Shawe-Taylor, J.; Platt, J.C.; others. Support vector method for novelty detection. NIPS. Citeseer, 1999, Vol. 12, pp. 582–588.
33. He, W.; Luo, Y.; Yin, F.; Hu, H.; Han, J.; Ding, E.; Liu, C.L. Context-aware mathematical expression recognition: An end-to-end framework and a benchmark. 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, December 4–8, 2016; Cancun, Mexico, pp. 3246–3251.
34. Gao, L.; Yi, X.; Liao, Y.; Jiang, Z.; Yan, Z.; Tang, Z. A deep learning-based formula detection method for PDF documents. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, November 10–15, 2017; Kyoto, Japan, Vol. 1, pp. 553–558.
35. Gao, L.; Yi, X.; Jiang, Z.; Hao, L.; Tang, Z. ICDAR2017 competition on page object detection. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, November 10–15, 2017; Kyoto, Japan, Vol. 1, pp. 1417–1422.
36. Yi, X.; Gao, L.; Liao, Y.; Zhang, X.; Liu, R.; Jiang, Z. CNN based page object detection in document images. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017, Vol. 1, pp. 230–235.
37. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.
38. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, June 18–22, 2018; Salt Lake City, Utah, USA, pp. 6154–6162.
39. Li, X.H.; Yin, F.; Liu, C.L. Page object detection from pdf document images by deep structured prediction and supervised clustering. 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, November 10–15, 2017; Kyoto, Japan, pp. 3627–3632.
40. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks, 2017, [arXiv:cs.CV/1703.06211].
41. Liu, Y.; Wang, Y.; Wang, S.; Liang, T.; Zhao, Q.; Tang, Z.; Ling, H. Cbnet: A novel composite backbone network architecture for object detection. Proceedings of the AAAI Conference on Artificial Intelligence, February 7–12, 2020; New York, USA, Vol. 34, pp. 11653–11660.
42. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **2012**, *25*, 1097–1105.

43. Siddiqui, S.A.; Malik, M.I.; Agne, S.; Dengel, A.; Ahmed, S. Decnt: Deep deformable cnn for table detection. *IEEE Access* **2018**, *6*, 74151–74161.
44. Agarwal, M.; Mondal, A.; Jawahar, C. CDeC-Net: Composite Deformable Cascade Network for Table Detection in Document Images. *arXiv preprint arXiv:2008.10831* **2020**.
45. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15–21, 2019; Long Beach, CA, USA, pp. 9308–9316.
46. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; others. Hybrid task cascade for instance segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 16–20, 2019; Long Beach, CA, USA, pp. 4974–4983.
47. Phong, B.H.; Hoang, T.M.; Le, T.L. A hybrid method for mathematical expression detection in scientific document images. *IEEE Access* **2020**, *8*, 83663–83684.
48. Chu, W.T.; Liu, F. Mathematical formula detection in heterogeneous document images. 2013 conference on technologies and applications of artificial intelligence. IEEE, December 6–8, 2013; Taipei, Taiwan, pp. 140–145.
49. Hashmi, K.A.; Liwicki, M.; Stricker, D.; Afzal, M.A.; Afzal, M.A.; Afzal, M.Z. Current Status and Performance Analysis of Table Recognition in Document Images with Deep Neural Networks. *IEEE Access* **2021**, *9*, 87663–87685.