Automatic content extraction in document images

Nicola Carkaxhija
Dept. of Information Engineering, University of Florence
Via di S. Marta 3, 50139 – Florence, Italy
nicola.carkaxhija@stud.unifi.it

Abstract

In this paper a revisited version of an automatic document content extraction method [1] is proposed. The algorithm relies on an image registration procedure, that in a first stage enables the definition of a probabilistic template from a set of scanned images, then it matches incoming images to the built model to locate the positions where variable fields appear. The approach is validated on collections of forms filled in different ways to observe how results vary with them.

1. Introduction

Despite the rise of the digital era, currently incoming data format of several entities still consists in paper. Migration of information stored in physical paper towards an electronic environment presents several advantages: storage, management, consultation, remote accessibility and security. Manual processing is a costly task and the presented method aims to make it automatic, eliminating the need of human intervention; this is particularly useful in historical collections where the template of the form is usually not available.

Information extraction is enabled by defining a template that predisposes documents to be read by an Optical Character Recognition (OCR) program, allowing to point out locations to be interpreted to extract the filled fields: indeed, the knowledge of the static part of a document allows, by contrast, to isolate the relevant content, that is represented by those portions that change over the instances. Thus, given a sufficiently large set of samples belonging to the same class, a fully automated process can generate a probabilistic template indicating the likelihood a pixel has to be either static or variable content.

Since the model is generated as average of many filled forms, with respect to it pixels in those parts that have been filled represent noise and thus may be interpreted as aleatory variables with a certain average removable by applying an opportune threshold.

In a summary, the method proposed in this paper consists of two different stages: in the former instances of the same document kind are provided to the system to generate a template, while in the latter the built model is used to process new incoming form images in order to extract the filled parts. Both the steps rests on an alignment algorithm illustrated hereinafter, that constitutes the main difference with respect to the original implementation to which the rest of the procedure abides.

2. Image registration

Given two instances belonging to the same class, image alignment is the task of finding the parameters that warp an image so that it minimizes the difference with the other one according a certain distance measure. In the case of a set with higher cardinality, it results in the representation of all the images into the same coordinate system.

To carry out this operation the homography estimation has been used instead of computing the Lukas and Kanade optical flow [2] as proposed in the original implementation of this content separation method, whereby several pre-processing steps are needed before alignment to make the algorithm attain more robustness, since otherwise it's sensible to local minima. Thus, the following operations became superfluous:

- connected component analysis to cut out irrelevant small regions too big ones that likely constitute margins of another document;

- the gradient module over the image to reduce the harmful contribution provided by too large foreground regions in the parameter estimation, potential cause of misalignments;

- the consequent binarization with the Otsu algorithm [3];

- the typical multiscale approach used to face out the algorithm sensitivity to local minima and avoid to get stuck before obtaining the actual wrapping parameters.

Furthermore, the use of homographies as wrapping functions makes possible to cope with perspective distortions introduced when the scanner cover doesn't completely lower, entailing a non-uniform pressure on the document during the acquisition, rather than handling just images related at most by an affine transform: scaling, translation, skewing and rotation.

With regard to the error to minimize, a much more robust method as RANSAC has been adopted in lieu of the squared sum to cope with the presence of outliers deriving as a side effect of the removal of the many expedients above-mentioned.

When loading the images, they are still processed using the adaptive binarization algorithm proposed by Sauvola [4] to lessen the effects of an uneven illumination and reduce amount of points involved in the computations.

3. Experimental results

The method has been tested on three document collections: the public NIST Structured Tax Forms Dataset (SPDB2) [5], whose fields has been filled through typewriters, a set of invoices with handwritten content and one of printed bills.

Many tries have been performed with sets of increasing size, but regardless the dataset involved, it could be claimed that a few dozen of samples are enough to produce a good template and the use of additional ones doesn't lead to perceptible improvements.

The best results have been obtained on the hand-filled invoices, foreseeable since such a content is characterized by a high variability.

The procedure yields good result also on NIST; however, not all the extracted fields are perfectly readable: specifically, this looks to be related to the use of a typewriter, that accurately centered at the begin of the blank space it's going to press on makes the first character camouflage as part of the template; because of the incline of the paper with respect to the machine this effect reduces as a long line is written.

In both the previous cases, content overlapping the template or too close to it is inevitably erased in the extraction phase.

Instead, print-filled invoices characters are always located in fixed position and assimilated by the template during its generation, so that the extraction fails and the content results in a few small sparse spots. In case of thin printed content printing, such a problem could be partially overcome by selecting images randomly, in order to avoid influences by document acquisition order or nomenclature: e.g. in consecutive documents, a field assuming the same value may be interpreted as part of the template.

However, while dealing with different datasets, it emerged that the generation of an almost perfect model is possible just by a regularizing intervention, therefore the method is not completely automatic: when increasing the contrast by applying a sigmoid over the image is necessary to adjust the grey-level intensity on which it is centered; to make the procedure really unsupervised such a value must be made adaptive. For this purpose, an Otsu binarization has been performed, but it brought as a side-effect a slight degradation in the extraction result.

4. Conclusions

By a qualitative analysis, the results achieved are satisfying, despite of some eroded characters that most of the times still are perceptible; in no way the method can suit to print-filled documents because unable to distinguish template and content.

The choice to don't use the Lukas and Kanade's algorithm in favour of a shorter and then simpler implementation worked, but it wasn't enough to reach a good result: there has been the need to carry out a further study to make the procedure effectively automatic.

References

[1] D. Aldavert, M. Rusinol and R. Toledo, *"Automatic Static/Variable Content Separation in Adiministrative Document Images"*, in 2017 14th IAPR International Conference on Document Analysis and Recognition.

[2] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, 1981, pp. 674–679.

[3] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, January 1979.

[4] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," ¨*Pattern Recognition*, vol. 33, no. 2, pp. 225–236.

[5] D. Dimmick, M. Garris, and C. L. Wilson, "Structured forms database," National Institutte of Standards and Technology.

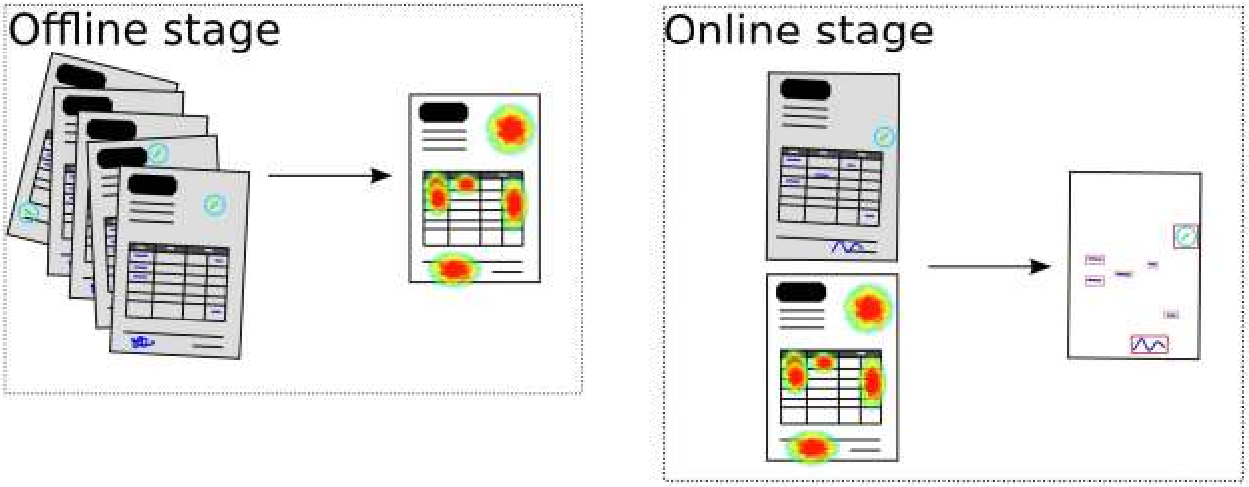


Figure 1: System overview, consisting in two phases: template generation and content extraction.

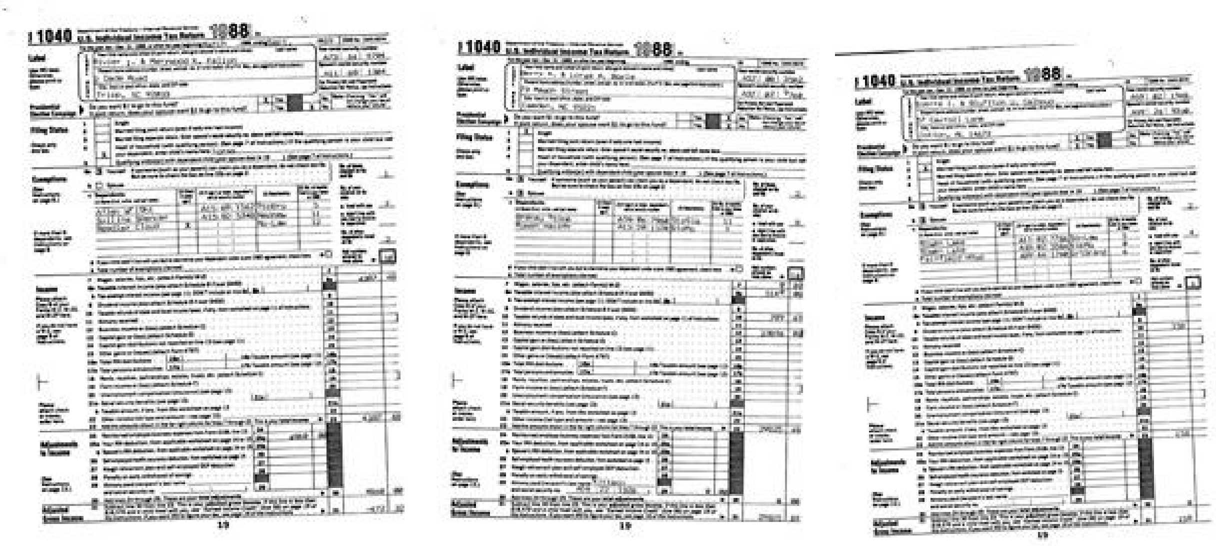
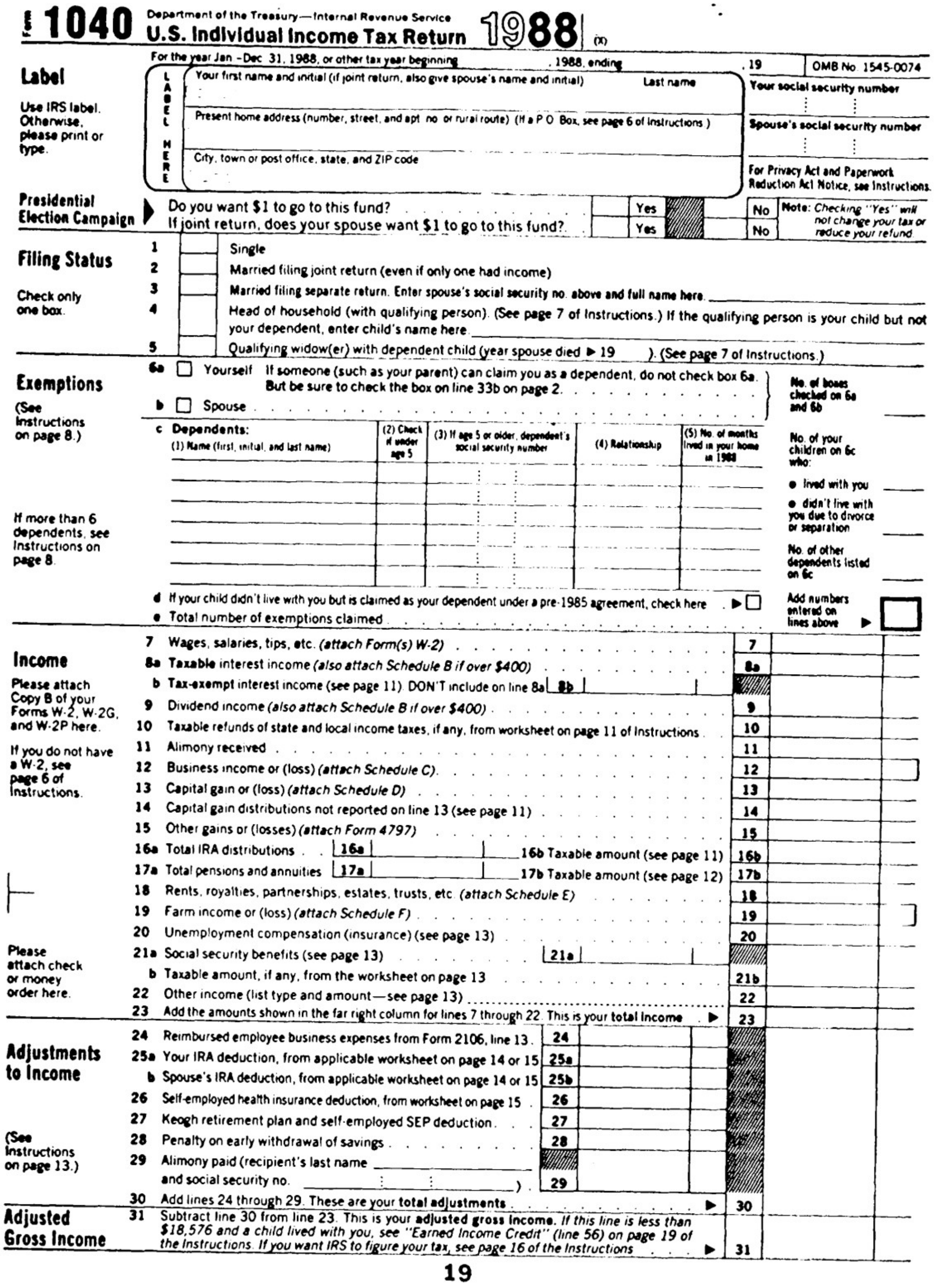


Figure 2: Samples of filled documents (with the same format) used to build the template.



Form **1040** Department of the Treasury—Internal Revenue Service
U.S. Individual Income Tax Return **1988** (X)

For the year Jan.–Dec. 31, 1988, or other tax year beginning , 1988, ending , 19 OMB No. 1545-0074

Label
Use IRS label. Otherwise, please print or type.

LABEL HERE
Your first name and initial (if joint return, also give spouse's name and initial) Last name
Present home address (number, street, and apt. no. or rural route) (If a P.O. Box, see page 6 of Instructions.)
City, town or post office, state, and ZIP code

Your social security number
Spouse's social security number
For Privacy Act and Paperwork Reduction Act Notice, see Instructions.

Presidential Election Campaign
Do you want $1 to go to this fund? Yes No
If joint return, does your spouse want $1 to go to this fund? Yes No
Note: *Checking "Yes" will not change your tax or reduce your refund.*

Filing Status
Check only one box.
1 □ Single
2 □ Married filing joint return (even if only one had income)
3 □ Married filing separate return. Enter spouse's social security no. above and full name here. ____
4 □ Head of household (with qualifying person). (See page 7 of Instructions.) If the qualifying person is your child but not your dependent, enter child's name here. ____
5 □ Qualifying widow(er) with dependent child (year spouse died ▶ 19). (See page 7 of Instructions.)

Exemptions
(See Instructions on page 8.)
6a □ Yourself If someone (such as your parent) can claim you as a dependent, do not check box 6a. But be sure to check the box on line 33b on page 2.
b □ Spouse
No. of boxes checked on 6a and 6b ____

c **Dependents:**

(1) Name (first, initial, and last name)	(2) Check if under age 5	(3) If age 5 or older, dependent's social security number	(4) Relationship	(5) No. of months lived in your home in 1988

If more than 6 dependents, see Instructions on page 8.

No. of your children on 6c who:
● lived with you ____
● didn't live with you due to divorce or separation ____
No. of other dependents listed on 6c ____

d If your child didn't live with you but is claimed as your dependent under a pre-1985 agreement, check here ▶ □
e Total number of exemptions claimed
Add numbers entered on lines above ▶ □

Income
Please attach Copy B of your Forms W-2, W-2G, and W-2P here.
If you do not have a W-2, see page 6 of Instructions.
Please attach check or money order here.

7 Wages, salaries, tips, etc. *(attach Form(s) W-2)* 7
8a **Taxable** interest income *(also attach Schedule B if over $400)* 8a
b **Tax-exempt** interest income (see page 11). DON'T include on line 8a 8b
9 Dividend income *(also attach Schedule B if over $400)* 9
10 Taxable refunds of state and local income taxes, if any, from worksheet on page 11 of Instructions 10
11 Alimony received 11
12 Business income or (loss) *(attach Schedule C)* 12
13 Capital gain or (loss) *(attach Schedule D)* 13
14 Capital gain distributions not reported on line 13 (see page 11) 14
15 Other gains or (losses) *(attach Form 4797)* 15
16a Total IRA distributions 16a 16b Taxable amount (see page 11) 16b
17a Total pensions and annuities 17a 17b Taxable amount (see page 12) 17b
18 Rents, royalties, partnerships, estates, trusts, etc. *(attach Schedule E)* 18
19 Farm income or (loss) *(attach Schedule F)* 19
20 Unemployment compensation (insurance) (see page 13) 20
21a Social security benefits (see page 13) 21a
b Taxable amount, if any, from the worksheet on page 13 21b
22 Other income (list type and amount—see page 13) 22
23 Add the amounts shown in the far right column for lines 7 through 22. This is your **total income** ▶ 23

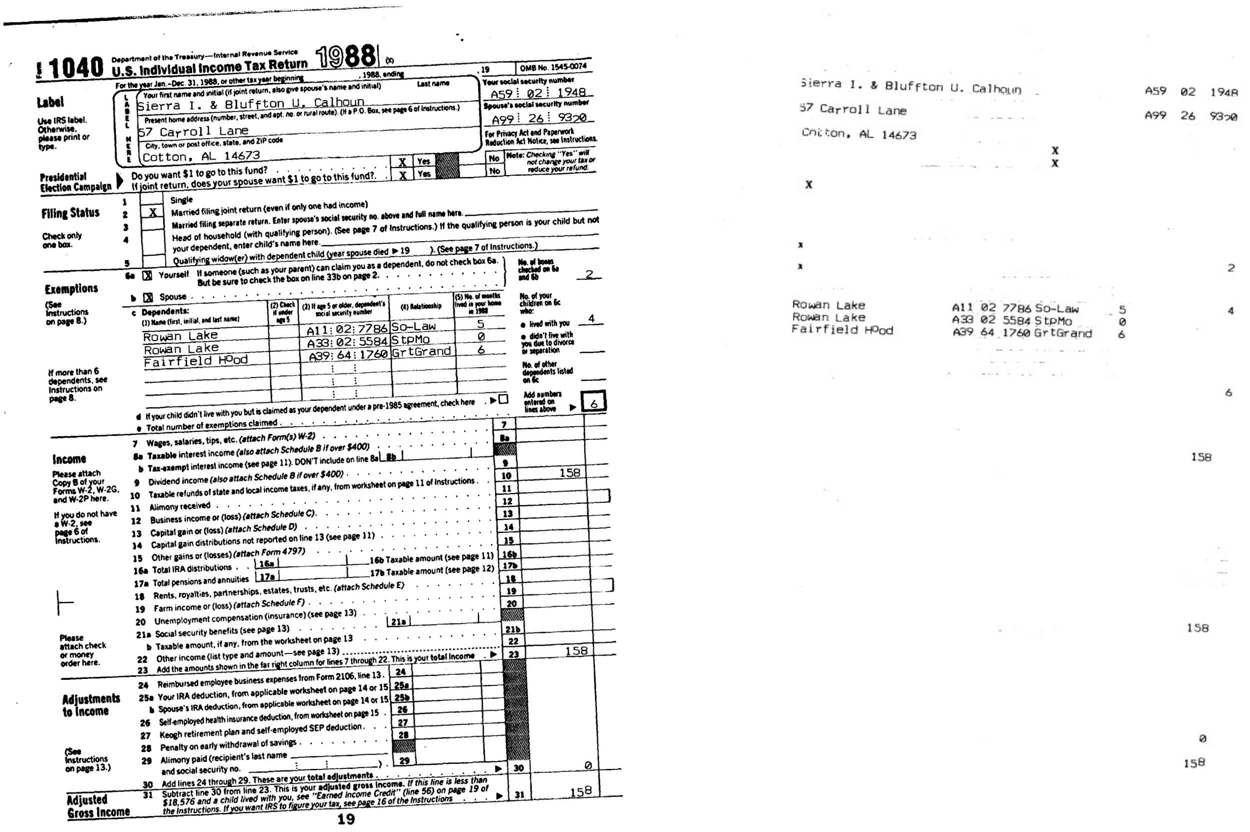
Adjustments to Income
(See Instructions on page 13.)

24 Reimbursed employee business expenses from Form 2106, line 13 24
25a Your IRA deduction, from applicable worksheet on page 14 or 15 25a
b Spouse's IRA deduction, from applicable worksheet on page 14 or 15 25b
26 Self-employed health insurance deduction, from worksheet on page 15 26
27 Keogh retirement plan and self-employed SEP deduction 27
28 Penalty on early withdrawal of savings 28
29 Alimony paid (recipient's last name ____ and social security no. ____) 29
30 Add lines 24 through 29. These are your **total adjustments** ▶ 30

Adjusted Gross Income
31 Subtract line 30 from line 23. This is your **adjusted gross income.** *If this line is less than $18,576 and a child lived with you, see "Earned Income Credit" (line 56) on page 19 of the Instructions. If you want IRS to figure your tax, see page 16 of the Instructions* ▶ 31

19

Figure 3: Template generated using 100 samples.



Form **1040** Department of the Treasury—Internal Revenue Service
U.S. Individual Income Tax Return **1988** (0)

For the year Jan.–Dec. 31, 1988, or other tax year beginning , 1988, ending , 19 | OMB No. 1545-0074

Label
Use IRS label. Otherwise, please print or type.
LABEL HERE

Your first name and initial (if joint return, also give spouse's name and initial) Last name
Sierra I. & Bluffton U. Calhoun
Present home address (number, street, and apt. no. or rural route). (If a P.O. Box, see page 6 of Instructions.)
57 Carroll Lane
City, town or post office, state, and ZIP code
Cotton, AL 14673

Your social security number
A59 02 1948
Spouse's social security number
A99 26 9320
For Privacy Act and Paperwork Reduction Act Notice, see Instructions.

Presidential Election Campaign ▶
Do you want $1 to go to this fund? X Yes No
If joint return, does your spouse want $1 to go to this fund?. X Yes No
Note: *Checking "Yes" will not change your tax or reduce your refund.*

Filing Status
Check only one box.
1 □ Single
2 X Married filing joint return (even if only one had income)
3 □ Married filing separate return. Enter spouse's social security no. above and full name here.
4 □ Head of household (with qualifying person). (See page 7 of Instructions.) If the qualifying person is your child but not your dependent, enter child's name here.
5 □ Qualifying widow(er) with dependent child (year spouse died ▶ 19). (See page 7 of Instructions.)

Exemptions
(See Instructions on page 8.)
6a ☒ Yourself If someone (such as your parent) can claim you as a dependent, do not check box 6a. But be sure to check the box on line 33b on page 2.
b ☒ Spouse
No. of boxes checked on 6a and 6b 2

c Dependents:

(1) Name (first, initial, and last name)	(2) Check if under age 5	(3) If age 5 or older, dependent's social security number	(4) Relationship	(5) No. of months lived in your home in 1988
Rowan Lake		A11: 02: 7786	So-Law	5
Rowan Lake		A33: 02: 5584	StpMo	0
Fairfield HOod		A39: 64: 1760	GrtGrand	6

If more than 6 dependents, see Instructions on page 8.

No. of your children on 6c who:
- lived with you 4
- didn't live with you due to divorce or separation
No. of other dependents listed on 6c
Add numbers entered on lines above ▶ 6

d If your child didn't live with you but is claimed as your dependent under a pre-1985 agreement, check here ▶ □
e Total number of exemptions claimed

Income
Please attach Copy B of your Forms W-2, W-2G, and W-2P here.
If you do not have a W-2, see page 6 of Instructions.
Please attach check or money order here.

7	Wages, salaries, tips, etc. *(attach Form(s) W-2)*	7	
8a	Taxable interest income *(also attach Schedule B if over $400)*	8a	
b	Tax-exempt interest income (see page 11). DON'T include on line 8a 8b		
9	Dividend income *(also attach Schedule B if over $400)*	9	
10	Taxable refunds of state and local income taxes, if any, from worksheet on page 11 of Instructions	10	158
11	Alimony received	11	
12	Business income or (loss) *(attach Schedule C)*	12	
13	Capital gain or (loss) *(attach Schedule D)*	13	
14	Capital gain distributions not reported on line 13 (see page 11)	14	
15	Other gains or (losses) *(attach Form 4797)*	15	
16a	Total IRA distributions 16a	16b Taxable amount (see page 11)	16b	
17a	Total pensions and annuities 17a	17b Taxable amount (see page 12)	17b	
18	Rents, royalties, partnerships, estates, trusts, etc. *(attach Schedule E)*	18	
19	Farm income or (loss) *(attach Schedule F)*	19	
20	Unemployment compensation (insurance) (see page 13)	20	
21a	Social security benefits (see page 13) 21a		
b	Taxable amount, if any, from the worksheet on page 13	21b	
22	Other income (list type and amount—see page 13)	22	
23	Add the amounts shown in the far right column for lines 7 through 22. This is your **total income** ▶	23	158

Adjustments to Income
(See Instructions on page 13.)

24	Reimbursed employee business expenses from Form 2106, line 13	24	
25a	Your IRA deduction, from applicable worksheet on page 14 or 15	25a	
b	Spouse's IRA deduction, from applicable worksheet on page 14 or 15	25b	
26	Self-employed health insurance deduction, from worksheet on page 15	26	
27	Keogh retirement plan and self-employed SEP deduction	27	
28	Penalty on early withdrawal of savings	28	
29	Alimony paid (recipient's last name and social security no.)	29	
30	Add lines 24 through 29. These are your **total adjustments** ▶	30	0

Adjusted Gross Income
31 Subtract line 30 from line 23. This is your **adjusted gross income.** *If this line is less than $18,576 and a child lived with you, see "Earned Income Credit" (line 56) on page 19 of the Instructions. If you want IRS to figure your tax, see page 16 of the Instructions* ▶ 31 158

19

```
Sierra I. & Bluffton U. Calhoun                A59  02  1948
57 Carroll Lane                                A99  26  9320
Cotton, AL 14673
                            X
                            X
 X


x

x                                                          2


Rowan Lake        A11 02 7786 So-Law     5                 4
Rowan Lake        A33 02 5584 StpMo      0
Fairfield HOod    A39 64 1760 GrtGrand   6



                                                           6



                                                         158




                                                         158




                                                           0
                                                         158
```

Figure 4: Test image and result of the extraction phase.