Reconocimiento Visual con Deep Learning: Detección de Objetos para el Reconocimiento de Montos

Canales G. Paula A.1 y Neira R. Alvaro R.2

*Departamento de Ciencias de la Computación  
Universidad de Chile*

1paulacanales@ug.uchile.cl

2alvaroneirareyes@gmail.com

***Abstract*— This document gives formatting instructions for authors preparing papers for publication in the Proceedings of an IEEE conference. The authors must follow the instructions given in the document for the papers to be published. You can use this document as both an instruction set and as a template into which you can type your own text.**

***Keywords***— **Include at least 5 keywords or phrases**

1. Introducción

El siguiente documento describe la tarea realizada para el curso Reconocimiento Visual con Deep Learning, el cual corresponde al diplomado de Inteligencia Artificial 2021.

El objetivo es entender, en forma práctica, el funcionamiento de detectores de objetos en imágenes, basados en modelos convolucionales, a través de una aplicación para el reconocimiento de montos manuscritos.

Se realizará el entrenamiento de un modelo de detección de objetos YOLO v5 para predecir el monto que aparece en una imagen de entrada, obtenida desde el dataset orand-car-with-bbs. Por último, se discutirán los resultados tanto en el conjunto de validación como en el conjunto de test.

Para el desarrollo de este trabajo, se utilizó un *notebook* en Google Colab [4] en el cual está disponible el código fuente y la visualización de resultados.

1. Desarrollo

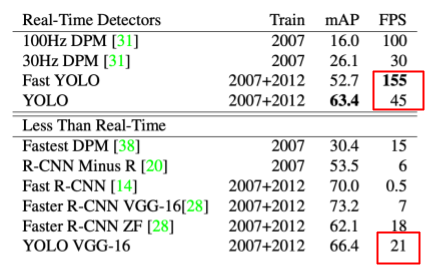
La detección de objetos consiste en localizar objetos de interés sobre una imagen. La localización consiste generalmente en predecir un rectángulo (a.k.a. bounding box ) definido por un punto de referencia (x,y) y el ancho y alto del rectángulo. Así, no debería ser difícil darnos cuenta que el reconocimiento de montos puede reducirse a un problema de detección de objetos, en donde los objetos de interés son los dígitos.

1. *Elección de Modelo de Detección*

Para resolver este problema se disponía de los siguientes modelos:

* RetinaNet.
* TridentNet.
* FCOS.
* YOLO versión 2 o superior.

TABLA I

Comparación FPS de primera versión de YOLO (extraída directo de [1])  


Se eligió YOLO (versión 5) por sobre las demás opciones por las siguientes razones:

* Velocidad de ejecución: Desde su introducción ([1]), las mediciones publicadas indican que su velocidad de detección es superior al de otros algoritmos similares en el año en que se probó. O por lo menos, otorga un buen *trade-off* entre precisión y eficiencia. Las tablas I y II indican estas comparaciones desde la fuente original. Estas corresponden a la versión del modelo 1 y 2 respectivamente.
* Debido a esta mayor eficiencia, nos fue posible probar más combinaciones de hiperparámetros y así poder optimizar aún más.
* Se eligió la versión 5 ([3]) porque fue probada con éxito en clases.

1. *Preparación de datos*

La implementación del modelo YOLO v5 requiere como primer paso, la conversión de las coordenadas de las anotaciones desde el formato *[class\_idx, xmin, ymin, width, height]* que viene con el dataset de origen, a un formato *[class\_idx, x\_center, y\_center, width, height]*, donde las coordenadas están normalizadas por el tamaño de la imagen.

Para esta conversión utilizamos el código empleado en clases [6] y el resultado es como el ejemplo detallado a continuación.

* Notación antigua

| 4: 83, 8, 43, 53 |
| --- |
| 6: 143, 11, 39, 50 |
| 0: 204, 11, 43, 43 |
| 6: 264, 15, 43, 43 |
| 6: 329, 11, 36, 51 |

* Notación nueva

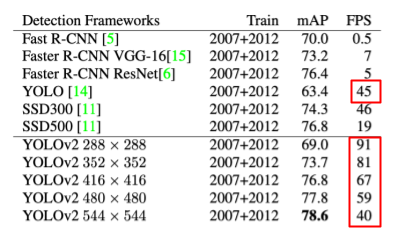
| 4 0.2301762114537445 0.4928571428571429 0.0947136563876652 0.7571428571428571 |
| --- |
| 6 0.3579295154185022 0.5142857142857142 0.08590308370044053 0.7142857142857143 |
| 0 0.4966960352422907 0.4642857142857143 0.0947136563876652 0.6142857142857143 |
| 6 0.6288546255506607 0.5214285714285715 0.0947136563876652 0.6142857142857143 |
| 6 0.76431718061674 0.5214285714285715 0.07929515418502203 0.7285714285714285 |

Finalmente, se realizó la división del dataset de origen en un conjunto de entrenamiento y uno de validación, con proporción 90% - 10% respectivamente. Por lo tanto, del total de 6506 imágenes, 5855 se utilizarán para el entrenamiento y 651 para la validación.

1. *Entrenamiento*

Se utilizan mayormente los pasos indicados en el *Notebook* *Colab* visto en clases ([5]).

TABLA II

Comparación FPS de YOLO9000 (extraída directo de [2])  


Se utiliza YOLOv5 disponible en GitHub [3]. Esta implementación requiere que las imágenes estén en una carpeta *images* y las anotaciones en texto, en la carpeta *labels*. Entonces el primer paso es modificar la estructura de archivos.

Junto con lo anterior, se escalan las coordenadas al rango 0.0-1.0, porque así lo requiere nuestro modelo elegido.

Se divide el conjunto de entrenamiento orand-car-with-bbs de forma 90%/10%. Es decir, 5855 imágenes para entrenar y 651 para hacer el testing.

Es necesario crear un nuevo archivo de configuración digits.yaml:

| #digits.yml  **train: /content/data/orand-car-with-bbs/training/train.txt**  **val: /content/data/orand-car-with-bbs/training/test.txt**  # number of classes  **nc: 10**  # class names  **names: ['0', '1', '2', '3', '4', '5', '6', '7', '8', '9']** |
| --- |

El entrenamiento propiamente tal requiere solo 1 linea *Colab* y requiere los siguientes parámetros:

* *--conf-thres (confidence threshold)*: Nivel de confianza que se requiere para las detecciones.
* *--iou-thres (NMS IoU threshold)*: Intersección sobre la unión, filtro de detecciones para un mismo objeto.

| !python train.py --img 640 --batch 16 --epochs 10 --data digits.yaml --weights yolov5s.pt |
| --- |

1. *Testing*

La detección sobre el conjunto de test se realizó ejecutando el archivo detect.py del repositorio de YOLO v5 utilizando los siguientes parámetros:

* *--conf-thres (confidence threshold)*: Nivel de confianza que se requiere para las detecciones.
* *--iou-thres (NMS IoU threshold)*: Intersección sobre la unión, filtro de detecciones para un mismo objeto.

El nivel de confianza utilizado fue de 0.5 y NMS de 0.1. Para llegar a estos valores se realizaron iteraciones probando diferentes valores y estos fueron los que entregan mejores resultados.

La detección genera una carpeta *detect* la cual contiene las imágenes del conjunto de test etiquetadas con las predicciones y además archivos de texto asociados a cada imágen con la información de las predicciones.

La información de las predicciones se muestran en el siguiente formato:

**<clase> <x>, <y>, <w>, <h>** donde:

* **clase**: es la clase del dígito predecida con valores que van desde 0 hasta 9.
* **x**: coordenada x de la esquina superior izquierda del bounding box.
* **y**: coordenada y de la esquina superior izquierda del bounding box.
* **w**: ancho del respectivo bounding box.
* **h**: alto respectivo bounding box.

1. *Accuracy conjunto de test*

Se implementó una clase Metrics en la cual existen dos funciones, una que realiza el cálculo de accuracy sobre el conjunto de test y otra que muestra los resultados negativos para facilitar el análisis de porqué el modelo falló en esos casos.

Para el cálculo de *accuracy*, se divide la cantidad de casos positivos en el total de casos. La contabilización de casos positivos se obtuvo comparando el monto detectado con el monto real, si estos coinciden entonces significa que es un caso positivo.

El monto real se obtuvo desde el archivo list.txt disponible en el dataset inicial, para simplificar su uso en el código, se realizó una copia de dicho archivo al repositorio de YOLO v5 en la carpeta runs. Luego, se utilizó un dataframe para poder identificar de mejor manera los datos del documento de texto.

[tabla]

El monto detectado se obtuvo de los archivos de texto generados en la detección, se revisó cada archivo de texto como dataframe, de tal manera que las clases se pudiesen ordenar respecto a la coordenada x para luego concatenar cada clase y formar el monto correspondiente.

La iteración se realizó por cada archivo de detección donde se comparan los montos y se almacenan los casos positivos y los negativos. Una vez finalizada la iteración se calcula el accuracy.

Para mostrar los resultados negativos, se recorre cada caso negativo y se genera una figura que muestra el manuscrito del monto correspondiente y como título se detalla el monto real y el monto detectado.

1. Resultados Experimentales y Discusión

A continuación se detallan los resultados obtenidos en el conjunto de validación entregados por YOLO v5, los resultados obtenidos en el conjunto de test calculados por nosotros y por último la discusión de lo anteriormente expuesto.

1. *Resultados conjunto de validación*

Se obtuvo un 97.168% de *precision* y 97.937% de *recall*, el comportamiento durante las 10 épocas se observa en Fig 1.

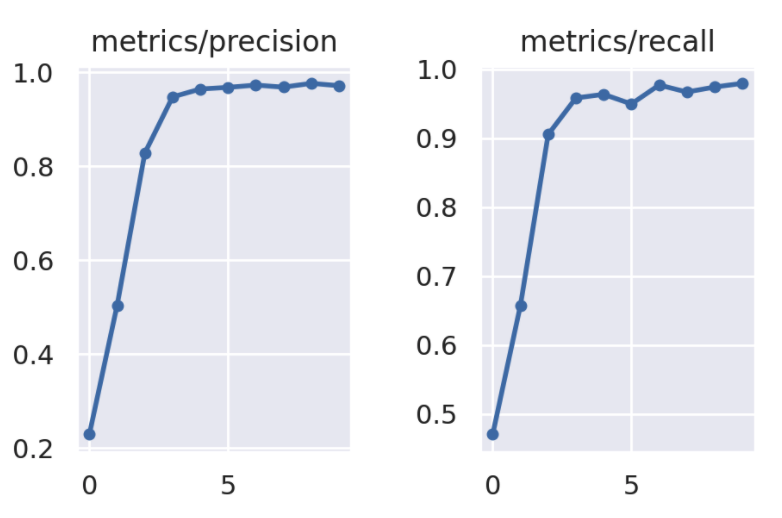


Fig. 1 Gráfico de métricas *precision* y *recall* durante 10 épocas del entrenamiento

Se observa que en términos de *precision*, con 5 épocas se alcanza un valor óptimo similar a las 10 épocas. Con respecto al *recall*, se considera que el valor alcanzado a las 10 épocas es mejor que si se hubieran utilizado 5 épocas. Por lo tanto, se definió que 10 épocas sería lo óptimo para la realización del entrenamiento.

Como sabemos, las métricas de *precision* y *recall* están relacionadas de manera directa y podemos visualizar su comportamiento en una curva, visualizada en la Fig.2.

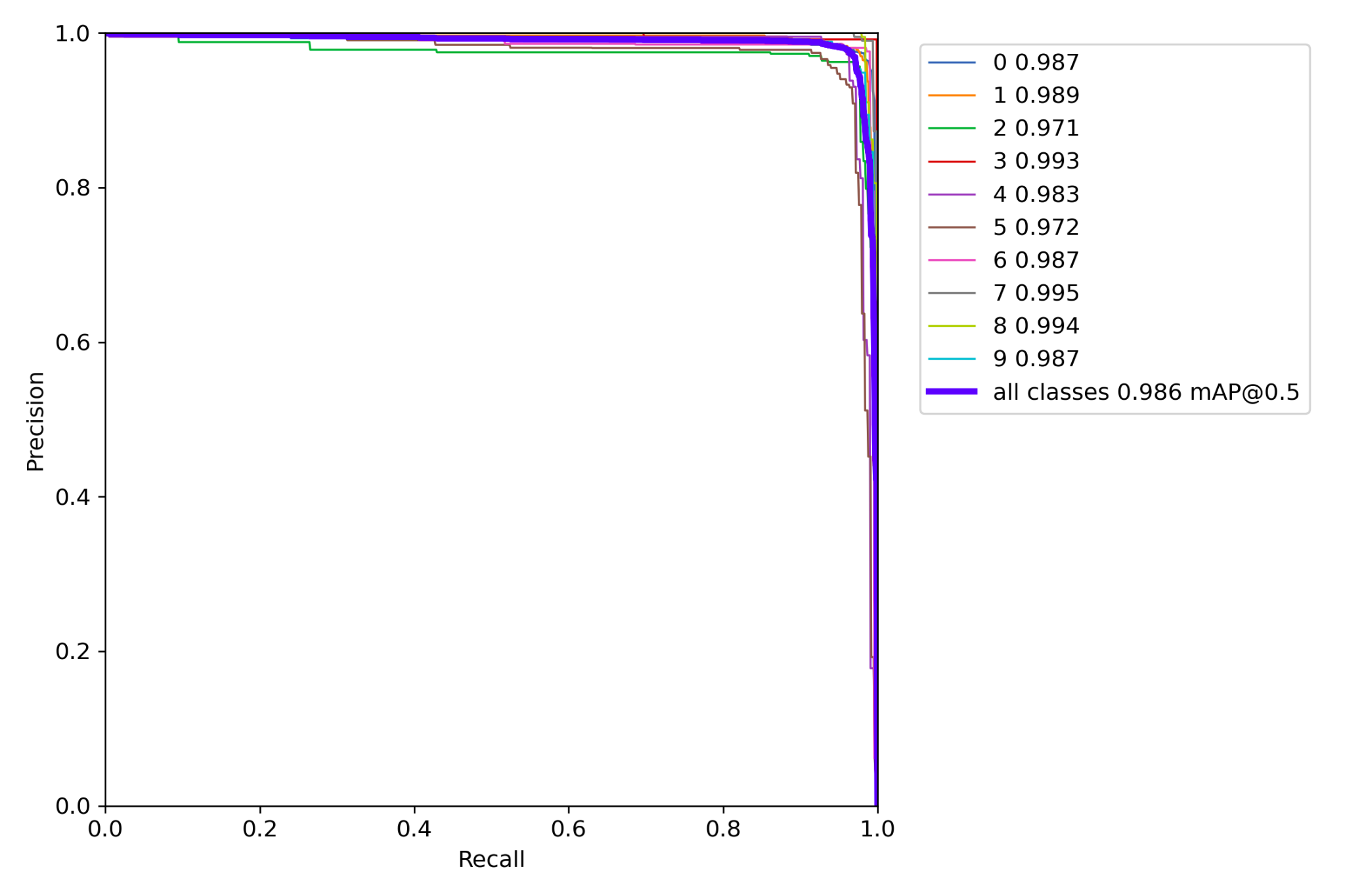


Fig. 2 Gráfico curva PR

Se observa que el rendimiento del modelo es bastante bueno ya que la curva se acerca notoriamente a la esquina superior derecha.

En la Fig. 2 se visualiza la matriz de confusión obtenida, la cual nos permite observar de forma explícita el comportamiento entre las clases.

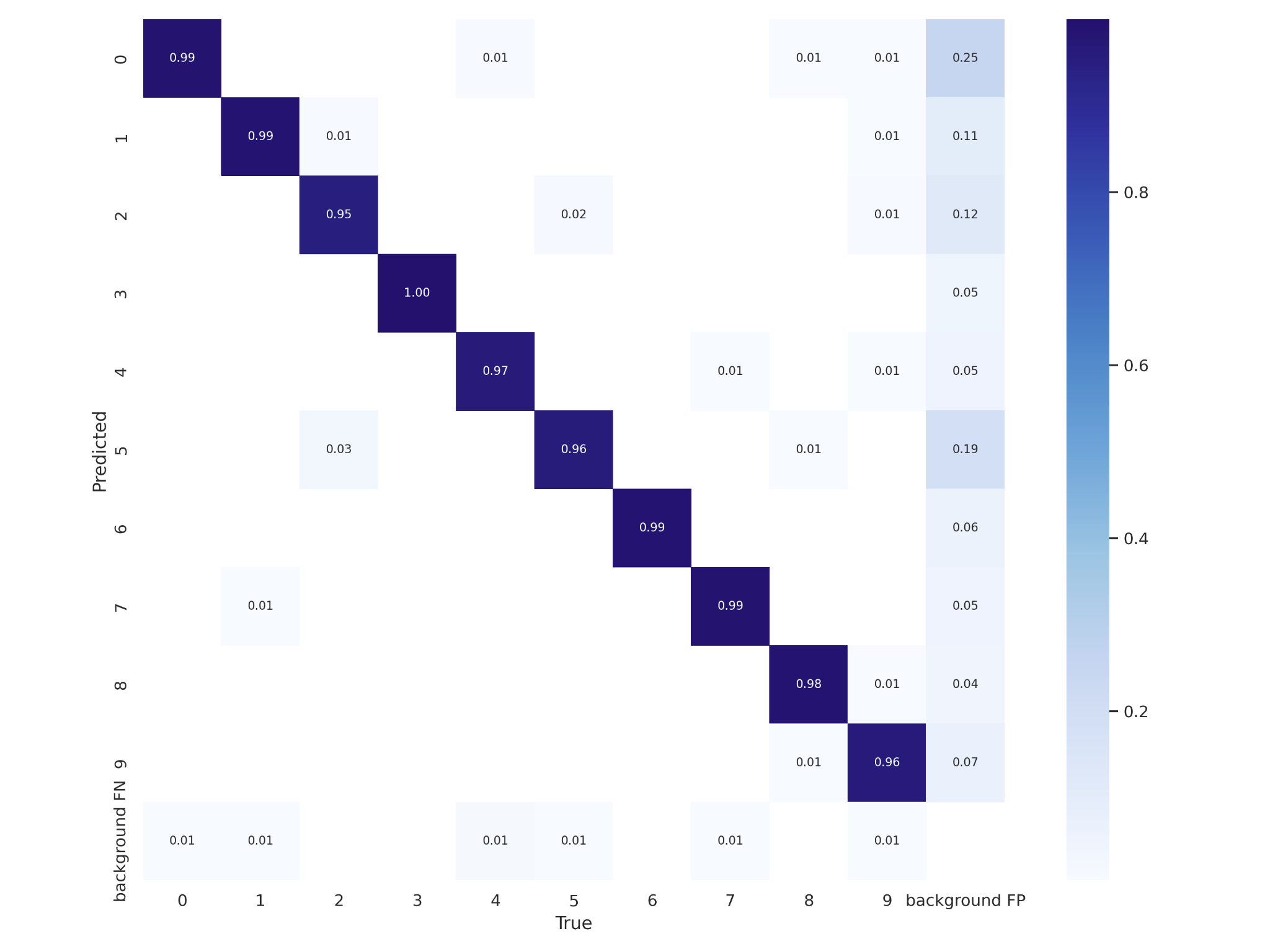


Fig. 1 Matriz de Confusión

En general se observa un buen comportamiento entre las clases, pero los dígitos que presentan más variabilidad en la clasificación serían el 2, 5 y 9. El dígito 2 suele confundirse con un 5 y 1, el dígito 5 con el 2 y el dígito 9 con el 0, 1, 2, 4 y 8.

El modelo de YOLO v5 genera las imágenes etiquetadas con cada predicción en el conjunto de validación, a continuación en la Fig. 3 se muestra un ejemplo de estas imágenes.

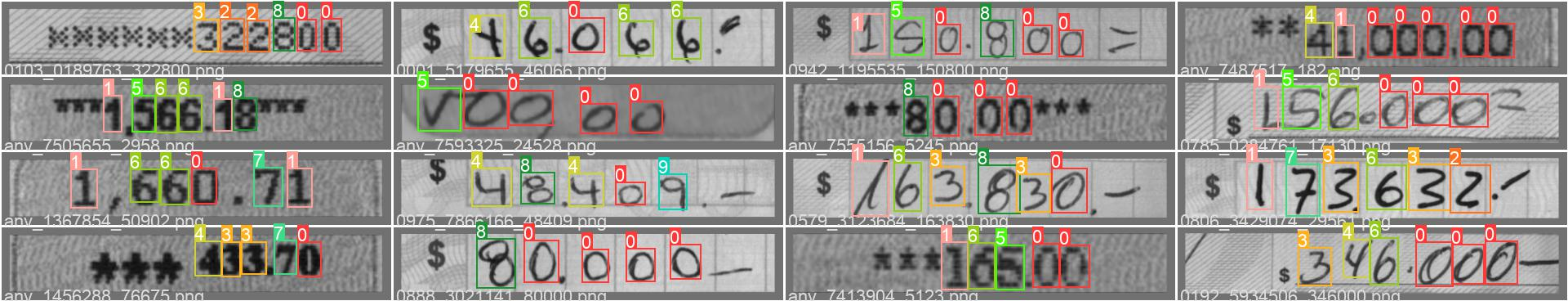


Fig. 1 Imágenes con etiquetas de montos detectados

Se observa que los montos predecidos acertaron en su totalidad.

1. *Resultados conjunto de test*

De un total de 463 archivos, 398 fueron casos positivos, por tanto se obtuvo un **85.96%** de *accuracy*.

Se identificó que los 65 casos negativos fueron mal detectados por diversas razones, las cuales se detallan a continuación con algunos ejemplos.

i. La mala calidad o poca visibilidad de imagen del monto manuscrito (Fig. 2)

|  |  |
| --- | --- |

Fig. 2 Montos mal detectados por poca visibilidad del monto real.

ii. La presencia de símbolos previo o posterior al monto manuscrito. (Fig. 2)

|  |  |
| --- | --- |

Fig. 2 Montos mal detectados por símbolos previo/posterior al monto.

iii. La forma de los números puede variar según la tipografía. (Fig. 2)

|  |  |
| --- | --- |

Fig. 2 Montos mal detectados por identificar el dígito 6 como 8 debido a la tipografía.

iv. Existe un caso en el que el monto real no corresponde a la imagen, fue mal etiquetado.

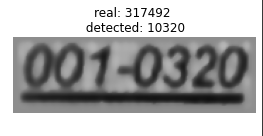


Fig. 2 Monto mal detectado por etiquetado incorrecto.

v. El modelo falla cuando el monto presenta gran cantidad de ceros correlativos, la detección omite uno o varios.

|  |  |
| --- | --- |

Fig. 2 Montos mal detectados por omisión de ceros.

vi. Hay nueve casos donde la detección entrega dos clases para un mismo dígito a pesar de haber utilizado un *iou-thres* en 0.1. Por ejemplo en la Fig. 2 podemos observar que el monto detectado 91332**25** y el 1229**25**1 fue erróneo ya que ese **25** destacado corresponde a dos clases, que puede ser un 2 o un 5. En todos los casos ocurre entre los dígitos 2 y 5.

|  |  |
| --- | --- |

Fig. 2 Montos mal detectados por intersección de clases.

La visualización de todos los casos negativos y el detalle de las coordenadas para los casos de falla por intersección se encuentran disponibles en el notebook de Google Colab [4].

1. Conclusiones

Detallar conclusiones encontradas

Deberíamos mencionar sobre los dígitos 2 y 5, lo del etiquetado en general

hay que decir que si existe relación entre los gráficos de las métricas con lo obtenido en el testing

References

1. Redmon, Joseph, Santosh Divvala, Ross Girshick, y Ali Farhadi. 2016. “You only look once: Unified, real-time object detection”. En Proceedings of the IEEE conference on computer vision and pattern recognition, 779–88. <https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html>.
2. Redmon, J., & Farhadi, A. (2016). *YOLO9000: Better, Faster, Stronger*.
3. Jocher, Glenn. 2020. “YOLOv5 GitHub”. GitHub. 29 de mayo de 2020. https://github.com/ultralytics/yolov5.
4. https://colab.research.google.com/drive/1OIwDmlyAmOia27brAc1nVt4wzS5uXWH6?usp=sharing (después arreglaré esta referencia)
5. Loyola. “Clase 6: detección de objetos”. [https://colab.research.google.com/drive](https://colab.research.google.com/drive/)/1tKjg4N\_jXidYcDdf8ee2--g\_hgQlD2yf

Recommended font sizes are shown in Table 1.

TABLE I  
Font Sizes for Papers

| **Font Size** | **Appearance (in Time New Roman or Times)** | | |
| --- | --- | --- | --- |
| **Regular** | **Bold** | **Italic** |
| 8 | table caption (in Small Caps),  figure caption,  reference item |  | reference item (partial) |
| 9 | author email address (in Courier),  cell in a table | abstract body | abstract heading (also in Bold) |
| 10 | level-1 heading (in Small Caps),  paragraph |  | level-2 heading,  level-3 heading,  author affiliation |
| 11 | author name |  |  |
| 24 | title |  |  |

All title and author details must be in single-column format and must be centered.

Every word in a title must be capitalized except for short minor words such as “a”, “an”, “and”, “as”, “at”, “by”, “for”, “from”, “if”, “in”, “into”, “on”, “or”, “of”, “the”, “to”, “with”.

Author details must not show any professional title (e.g. Managing Director), any academic title (e.g. Dr.) or any membership of any professional organization (e.g. Senior Member IEEE).

To avoid confusion, the family name must be written as the last part of each author name (e.g. John A.K. Smith).

Each affiliation must include, at the very least, the name of the company and the name of the country where the author is based (e.g. Causal Productions Pty Ltd, Australia).

Email address is compulsory for the corresponding author.

1. *Section Headings*

No more than 3 levels of headings should be used. All headings must be in 10pt font. Every word in a heading must be capitalized except for short minor words as listed in Section III-B.

1. *Level-1 Heading*: A level-1 heading must be in Small Caps, centered and numbered using uppercase Roman numerals. For example, see heading “III. Page Style” of this document. The two level-1 headings which must not be numbered are “Acknowledgment” and “References”.
2. *Level-2 Heading:* A level-2 heading must be in Italic, left-justified and numbered using an uppercase alphabetic letter followed by a period. For example, see heading “C. Section Headings” above.
3. *Level-3 Heading:* A level-3 heading must be indented, in Italic and numbered with an Arabic numeral followed by a right parenthesis. The level-3 heading must end with a colon. The body of the level-3 section immediately follows the level-3 heading in the same paragraph. For example, this paragraph begins with a level-3 heading.
4. *Figures and Tables*

Figures and tables must be centered in the column. Large figures and tables may span across both columns. Any table or figure that takes up more than 1 column width must be positioned either at the top or at the bottom of the page.

Graphics may be full color. All colors will be retained on the CDROM. Graphics must not use stipple fill patterns because they may not be reproduced properly. Please use only *SOLID FILL* colors which contrast well both on screen and on a black-and-white hardcopy, as shown in Fig. 1.

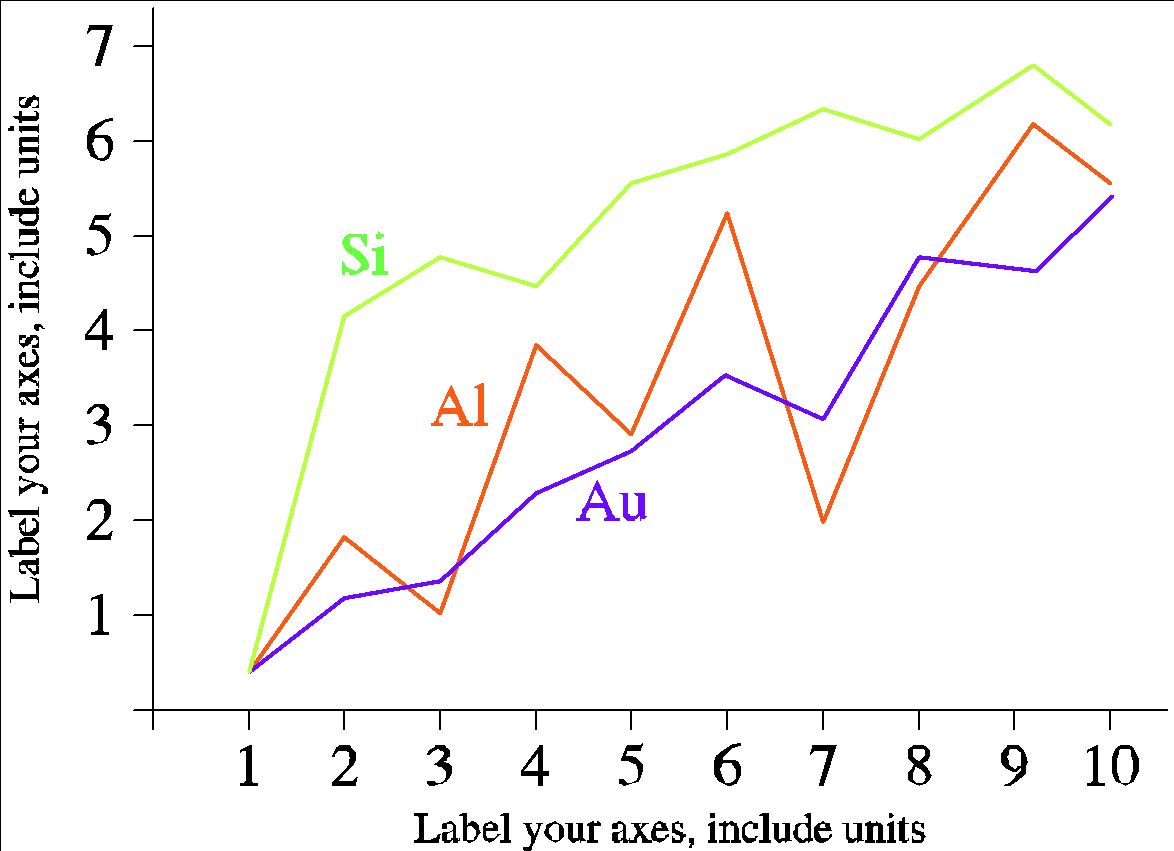


Fig. 1 A sample line graph using colors which contrast well both on screen and on a black-and-white hardcopy

Fig. 2 shows an example of a low-resolution image which would not be acceptable, whereas Fig. 3 shows an example of an image with adequate resolution. Check that the resolution is adequate to reveal the important detail in the figure.

Please check all figures in your paper both on screen and on a black-and-white hardcopy. When you check your paper on a black-and-white hardcopy, please ensure that:

* the colors used in each figure contrast well,
* the image used in each figure is clear,
* all text labels in each figure are legible.

1. *Figure Captions*

Figures must be numbered using Arabic numerals. Figure captions must be in 8 pt Regular font. Captions of a single line (e.g. Fig. 2) must be centered whereas multi-line captions must be justified (e.g. Fig. 1). Captions with figure numbers must be placed after their associated figures, as shown in Fig. 1.

[](http://www.freedigitalphotos.net/images/Television_Video_and_g178-TV_p11679.html)

Fig. 2 Example of an unacceptable low-resolution image

[](http://www.freedigitalphotos.net/images/Television_Video_and_g178-TV_p11678.html)

Fig. 3 Example of an image with acceptable resolution

1. *Table Captions*

Tables must be numbered using uppercase Roman numerals. Table captions must be centred and in 8 pt Regular font with Small Caps. Every word in a table caption must be capitalized except for short minor words as listed in Section III-B. Captions with table numbers must be placed before their associated tables, as shown in Table 1.

1. *Page Numbers, Headers and Footers*

Page numbers, headers and footers must not be used.

1. *Links and Bookmarks*

All hypertext links and section bookmarks will be removed from papers during the processing of papers for publication. If you need to refer to an Internet email address or URL in your paper, you must type out the address or URL fully in Regular font.

1. *References*

The heading of the References section must not be numbered. All reference items must be in 8 pt font. Please use Regular and Italic styles to distinguish different fields as shown in the References section. Number the reference items consecutively in square brackets (e.g. [1]).

When referring to a reference item, please simply use the reference number, as in [2]. Do not use “Ref. [3]” or “Reference [3]” except at the beginning of a sentence, e.g. “Reference [3] shows …”. Multiple references are each numbered with separate brackets (e.g. [2], [3], [4]–[6]).

Examples of reference items of different categories shown in the References section include:

* example of a book in [1]
* example of a book in a series in [2]
* example of a journal article in [3]
* example of a conference paper in [4]
* example of a patent in [5]
* example of a website in [6]
* example of a web page in [7]
* example of a databook as a manual in [8]
* example of a datasheet in [9]
* example of a master’s thesis in [10]
* example of a technical report in [11]
* example of a standard in [12]

1. Conclusions

The version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal Productions from the IEEE LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templates are self-contained. Causal Productions has used its best efforts to ensure that the templates have the same appearance.

Acknowledgment

The heading of the Acknowledgment section and the References section must not be numbered.

Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template. To see the list of contributors, please refer to the top of file IEEETran.cls in the IEEE LaTeX distribution.

References

1. S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
2. J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.
3. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT,” *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.
4. M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, “High resolution fiber distributed measurements with coherent OFDR,” in *Proc. ECOC’00*, 2000, paper 11.3.4, p. 109.
5. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digital-to-RF converter,” U.S. Patent 5 668 842, Sept. 16, 1997.
6. (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
7. M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/
8. *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.
9. “PDCA12-70 data sheet,” Opto Speed SA, Mezzovico, Switzerland.
10. A. Karnik, “Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP,” M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.
11. J. Padhye, V. Firoiu, and D. Towsley, “A stochastic model of TCP Reno congestion avoidance and control,” Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
12. *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.