

Software inteligente para la digitalización de placas espectroscópicas

Franco Ronchetti^{1,4}  Facundo Quiroga^{1,5}  Nehuén Pereyra¹ Joaquín Miranda¹ Santiago Ponte¹ Yael Aidelman^{2,3}  Roberto Gamen^{2,3}  and Laura Lanzarini¹ 

¹ Instituto de Investigación en Informática LIDI, Facultad de Informática, Universidad Nacional de La Plata (UNLP), La Plata, Argentina

² Facultad de Ciencias Astronómicas y Geofísicas, UNLP, La Plata, Argentina

³ Instituto de Astrofísica de La Plata, CONICET-UNLP, La Plata, Argentina

⁴ Comisión de Investigaciones Científicas de la Pcia. De Bs. As. (CIC-PBA), La Plata, Argentina

⁵ Becario Postdoctoral UNLP

{fronchetti, fquiroga}@lidi.info.unlp.edu.ar

Resumen La Facultad de Ciencias Astronómicas y Geofísicas (FCAG) de la Universidad Nacional de La Plata (UNLP) tiene un rico historial de observaciones astronómicas, geofísicas y meteorológicas. En particular, su colección de observaciones incluye 15.000 registros espectroscópicos en placas de vidrio obtenidas entre 1920 y 1980. La recuperación de los registros espectroscópicos implica un proceso complejo que incluye varias etapas: la recopilación de las placas fotográficas y sus respectivos metadatos; el escaneo de las placas y su conversión a archivos con formatos útiles a la astronomía y la extracción de los espectros y su calibración en longitud de onda.

Dada la cantidad de registros y las dificultades de su procesamiento, usando métodos manuales la recuperación del patrimonio completo llevaría una cantidad de tiempo prohibitiva y podría introducir diferencias en la sistematización y el análisis de los registros.

En este artículo presentamos un software inteligente para ayudar en la recuperación de la información contenida en las placas. El software contiene un módulo que detecta de forma automática los espectros, agiliza la carga de los metadatos y verifica su estandarización. El uso del software permitirá reducir significativamente las horas-persona necesarias y los errores de procesamiento.

Keywords: Espectros · YOLO · Placa espectroscópica · ReTrOH · Universidad Nacional de La Plata · Astronomía

1. Introducción

En el año 2000 la Unión Astronómica Internacional (IAU) emitió una resolución (Nro. B3 *Safeguarding the information in photographic plates*) [11], donde

solicita que se tomen medidas para conservar los datos históricos de todas las fuentes astronómicas, ya que, de no hacerlo, se perderán para las futuras generaciones de astrónomos. Agrega, además, que se debe procurar la transferencia de los datos históricos a medios modernos, los cuales deberán proveer su acceso a toda la comunidad internacional para el bien de toda la investigación astronómica. Desde ese entonces, en distintas partes del mundo se ha comenzado con la labor de digitalización de placas fotográficas, por ejemplo: la colección de placas astronómicas de Harvard⁶ o la colección del Observatorio Maria Mitchell⁷, por mencionar algunas. Particularmente en Argentina, por esos años, se comenzaron a digitalizar placas fotográficas creando el Primer archivo digital de placas fotográficas del Observatorio Astronómico de Córdoba [1].

El interés por la digitalización no es solo histórico; a partir de estos procesos de recuperación se han obtenido resultados importantes. Por ejemplo, se detectaron variaciones de largo período en blazares [13], se obtuvieron posiciones y magnitudes estelares [7], se detectaron y clasificaron nuevos objetos [5], se construyeron curvas de luz de estrellas T Tauri [4], se descubrieron eventos eruptivos en la estrella Variable Luminosa Azul R71 ocurridos a comienzos del siglo pasado y nunca reportados [12], se detectó un cúmulo abierto entre las estrellas de campo de la región de Collinder 132 y se calcularon el movimiento propio medio y las probabilidades de pertenencia de las estrellas de la región [8], por mencionar algunos.

La FCAG cuenta con una gran colección de placas espectrográficas y fotográficas en formato de vidrio (Figura 1). Estas observaciones fueron realizadas entre las décadas del '20 y del '80 por renombrados astrónomos argentinos como R. Barbá, V. Niemela, E. Brandi, L. García, O. Ferrer, N. Morrell, H. Levato, J. Sahade y A. Ringuelet entre otros. Se estima que en la FCAG hay más de 15.000 espectros registrados en placas, tomados con instrumentos instalados en el Observatorio de La Plata (OALP), el Observatorio Astronómico de Córdoba (OAC) y el Observatorio Interamericano de Cerro Tololo (CTIO) en Chile [3].

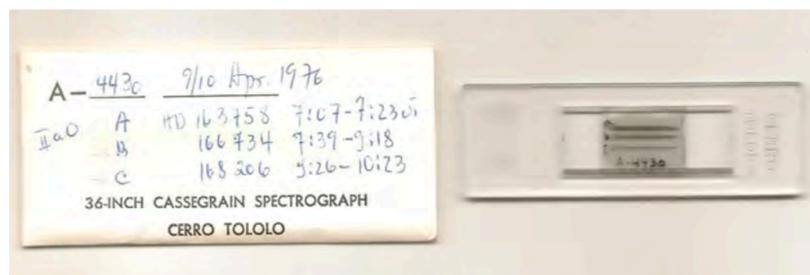


Figura 1. Placa espectroscópica, conteniendo tres espectrogramas, junto a sus anotaciones en formato papel.

⁶ <http://dasch.rc.fas.harvard.edu/project.php>

⁷ <https://www.mariamitchell.org/astronomical-plates-collection>

Sin embargo, la forma en que estos datos están disponibles no es útil, ya que no permite acceder a ellos con herramientas y software modernos. Además, el instrumental destinado a estos procesos ya no se encuentra operativo. Por esta razón, el 1 de mayo de 2019 se institucionalizó en la FCAG la creación de un repositorio científico para la preservación de este acervo histórico-científico-cultural mediante el llamado proyecto de Recuperación del Trabajo Observacional Histórico (ReTrOH)

Junto al Nuevo Observatorio Virtual Argentino (NOVA⁸), con sede en la FCAG, se ha comenzado el proceso de recuperación de datos históricos utilizando un escáner Nikon 9000ED. Hasta hoy se han digitalizado alrededor de 250 placas que se encuentran disponibles en el Repositorio Institucional del Servicio de Difusión de la Creación Intelectual (SeDiCI) de la UNLP. Se han analizado de forma manual dos espectros, que fueron procesados y analizados para validar la calidad de la recuperación de los mismos con fines científicos. Estos resultados dieron lugar a una tesis de licenciatura [6]. Si bien el proceso fue exitoso en cuanto a la validez del producto científico final, i.e. espectro digital, se concluyó que tal procedimiento no podría extrapolarse a las miles de placas existentes debido a la gran cantidad de tiempo necesario. Además, la naturaleza manual del procesamiento puede introducir errores sistemáticos o criterios subjetivos de procesamiento.

Por otro lado, los métodos de Inteligencia Artificial permiten automatizar varias de las tareas repetitivas comúnmente realizadas por humanos, particularmente aquellos basados en datos como el Aprendizaje Automático. Las Redes Neuronales son los modelos de aprendizaje automático con mejor desempeño en la actualidad en una gran variedad de problemas, particularmente en áreas de Visión por Computadora. En los últimos años, se ha conseguido entrenar Redes Neuronales con múltiples capas mediante un conjunto de técnicas que suelen denominarse Aprendizaje Profundo (*Deep Learning*) [2].

En este artículo presentamos un sistema que permite agilizar y automatizar el proceso completo de digitalización de estas placas. El objetivo del sistema es detectar los diferentes espectros que se encuentran en cada placa, recortarlos y exportarlos al formato FITS (por sus siglas en inglés *Flexible Image Transport System*), estándar en astronomía. La detección se realizó con un modelo de *Deep Learning* entrenado específicamente para el problema. El sistema además permite ingresar y verificar los metadatos asociados a cada espectro de la placa.

En el resto de esta sección, introducimos conceptos básicos sobre las placas espectrográficas. En la sección 1.1, detallamos los aspectos principales del desarrollo del software, dejando para la sección 3 los detalles de la detección automática de espectros. La sección 4 presenta las conclusiones y el trabajo futuro esperado.

⁸ <https://nova.conicet.gov.ar>

1.1. Placas Espectroscópicas

En tiempos previos a la era digital, se utilizaban diversos sistemas para registrar las observaciones. Los más comunes eran las placas fotográficas. Estas consisten de una base de vidrio con una emulsión fotosensible adherida a una de sus caras. Mediante efecto fotoquímico, la emulsión es capaz de reaccionar a los fotones, portadores de información astrofísica. La intensidad de la señal queda visibilizada en el oscurecimiento de la emulsión. La figura 1 muestra un ejemplo de estas placas junto con anotaciones manuscritas realizadas por el observador.

Las placas fotográficas se utilizaron desde fines del siglo XIX hasta bien entrada la década del '80 del siglo pasado y si bien actualmente se encuentran obsoletas, es importante remarcar su simplicidad y el gran tamaño que pueden llegar a alcanzar (característica ideal para realizar relevamientos de grandes áreas del cielo).

Generalmente, las placas se encuentran almacenadas dentro de sobres de papel. En su mayoría, los mismos cuentan con la información de la observación como el nombre del objeto observado, nomenclatura de placa, fecha y hora de observación, tipo de lámpara de comparación, tipo de emulsión, observador, observatorio e instrumental utilizado, etc.

Espectros El objetivo general de la digitalización de las placas consiste en tener los espectros de ciencia disponibles para la comunidad científica. La figura 2 muestra un ejemplo de un espectro digitalizado y recortado. Un espectro es una imagen en la cual se observa la cantidad de energía recibida por longitud de onda. La manera de obtener una imagen de este estilo consiste en hacer incidir la luz sobre un objeto dispersor, como por ejemplo un prisma o una red de difracción, logrando así descomponer la luz “blanca” en todos sus “colores” (diferentes longitudes de onda).

Este “arco iris” así formado, contiene información sobre el emisor de la luz, como su temperatura y los elementos químicos que lo componen, y se denomina espectro. Los espectros no son una mera secuencia de colores sino que en ellos se pueden identificar una serie de líneas oscuras superpuestas. Estas líneas son las que contienen la información sobre los elementos químicos presentes, ya que cada elemento de la tabla periódica tiene asociada una serie de líneas particulares y exclusivas (como las huellas digitales en los seres humanos). El espectro de una estrella trae consigo esas huellas.

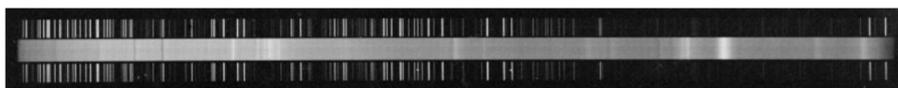


Figura 2. Espectro digitalizado. En la región central se observa el espectro de ciencia acompañado de los espectros de la lámpara de comparación (arriba y abajo).

Los espectros de los objetos de estudio adquiridos, llamados espectros de ciencia, siempre van acompañados de “espectros de comparación”. Estos últimos corresponden a un espectro de una lámpara con gases cuya secuencia de líneas es bien conocida (generalmente se utilizan lámparas de hierro, de neón y argón, o de helio-neón-argón, entre otras) lo que permite relacionar sus posiciones en la placa con la longitud de onda que les corresponde. Este proceso se denomina calibración en longitud de onda. En la figura 2 se puede observar un espectro estelar en la región central⁹, y arriba y abajo, los espectros de la lámpara de comparación (espectro de líneas brillantes).

2. Software desarrollado

El software desarrollado permite al equipo científico realizar las digitalizaciones de las placas de un modo más ágil y ordenado. Está disponible públicamente para su uso, y su desarrollo es bajo el modelo de software libre ¹⁰

El objetivo del sistema es que una vez generado un archivo imagen con la placa digitalizada, el sistema permita el recorte de los diversos espectros que haya y la carga de los metadatos de cada uno. Cada espectro debe ser exportado en formato FITS. El software posee los siguientes componentes:

- Frontend HTML/Javascript desarrollado con el framework Svelte que puede ejecutarse en cualquier navegador web. Este permite la definición de un directorio de trabajo donde es posible seleccionar la imagen a procesar. Se visualizan los diferentes espectros detectados en la imagen y permite la carga de los metadatos correspondientes a cada uno (ver figura 3).
- Backend Python desarrollado con el framework Flask que realiza el vínculo entre la detección de los espectros y la aplicación web.
- Detector de espectros (ver sección 3) que permite detectar automáticamente los espectros en la imagen digitalizada, con el consiguiente ahorro del tiempo de procesamiento de las imágenes.

En el frontend, el sistema permite la definición de un directorio de trabajo, donde se podrá elegir cada imagen a tratar. El sistema persiste el estado en el que se encuentra cada archivo del directorio: sin procesar, en curso, o terminado. Al estar en curso, es posible recargar la sesión de trabajo del archivo, junto con los espectros detectados y los metadatos que se hayan cargado de ellos. El sistema permite además calibrar el brillo, contraste y filtro de color de la imagen cargada, para su mejor manipulación. Estos cambios no se ven reflejados en el recorte final de cada objeto.

Un aspecto fundamental en el sistema es el manejo adecuado de los metadatos. Por un lado, los archivos FITS que se generan deben tener ciertos metadatos convenidos mundialmente [9], incluso si no se tiene información sobre estos para

⁹ En este caso particular se observa un espectro combinado, con líneas brillantes (más blancas) y líneas oscuras (más negras).

¹⁰ <https://github.com/midusi/spectrogram>

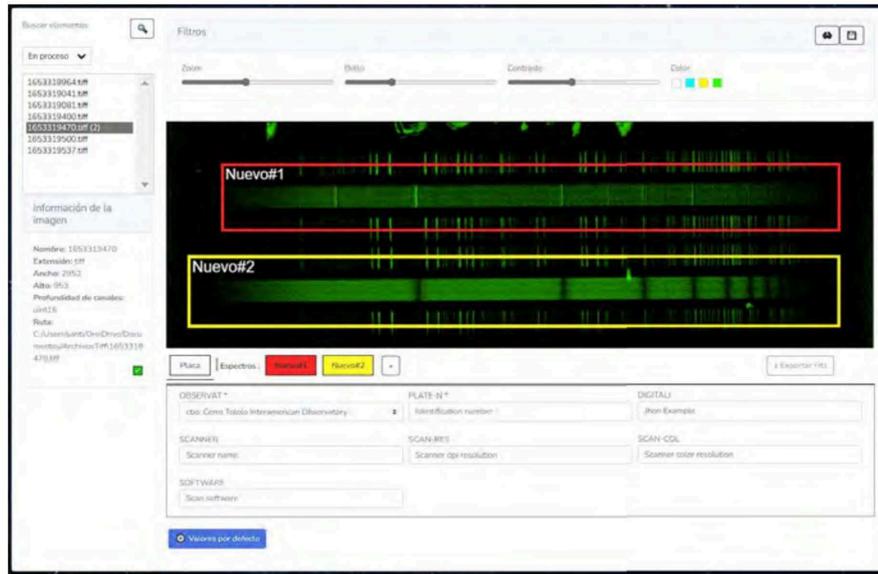


Figura 3. Frontend del sistema desarrollado. A la izquierda puede verse el directorio de trabajo e información del archivo. A la derecha puede verse la imagen escaneada junto con los espectros detectados y la carga de los metadatos.

un espectro particular. Por otro lado, el sistema permite diferenciar entre dos tipos de metadatos, los referentes a la placa donde estaba el espectro y los referentes al objeto observado. La información de la placa debe ser replicada en cada archivo FITS, ya que es común a todos los espectros (por ejemplo, número de placa, nombre del observador, lugar de observación, entre otros). Por ende, el sistema tiene facilidades para realizar esta carga de metadatos de forma completa de manera manual.

No obstante, para alivianar la carga, también permite buscar en forma automática en la base de datos SIMBAD¹¹ en el caso en que el objeto a cargar sea conocido previamente. Para esto, se debe indicar tres campos obligatorios: nombre del objeto, día de observación y la hora en formato universal. Con esta información el sistema puede completar, en la mayoría de los casos, una gran parte de los metadatos como el tipo de objeto y sus coordenadas.

3. Detección automática de Espectros

Con el fin de agilizar el trabajo del profesional que debe digitalizar las placas, desarrollamos un detector automático de espectros dentro de la placa. Para esto, confeccionamos una base de datos y entrenamos un modelo Convolutivo, como se menciona en las siguientes secciones.

¹¹ <http://simbad.cds.unistra.fr/simbad>

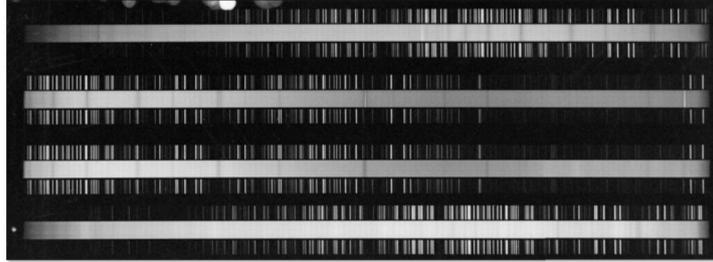


Figura 4. Ejemplo de placa digitalizada con cuatro espectros distintos.

3.1. Desafíos

Si bien la identificación manual resulta ser relativamente simple, existen diferentes desafíos a tener en cuenta para realizar la detección de forma automática, como se comentan a continuación:

- **Varios espectros por imagen.** Como puede verse en la figura 4, lo más usual es tener diversos espectros en una placa. Esto no sería un gran problema, pero sí es necesario abordar la estrategia como un problema de detección dentro de una imagen.
- **Placas deterioradas.** Se encontraron diversas placas con suciedad o manchas inherente al tiempo de almacenamiento o a diferentes problemas al realizar la observación. En la figura 5 pueden verse algunos ejemplos.
- **Falta de lámparas de comparación.** En algunas ocasiones el espectro carece de sus lámparas de comparación, o incluso existen lámparas sin el espectro correspondiente. Estas situaciones deben ser descartadas ya que no permite hacer un análisis de los datos de forma apropiada. La figura 5 muestra un ejemplo de esto. Estos ejemplos fueron caracterizados como *negativos* al momento de realizar la base de datos (ver sección siguiente).

3.2. Base de datos

En primer lugar, construimos una base de datos para el problema particular de la detección de los espectros, ya que no hemos encontrado una similar que pueda ser utilizada. Para esto, utilizamos 256 imágenes de placas digitalizadas donde cada una contiene un promedio de 3 espectros cada una, dando un total de aproximadamente 800 espectros etiquetados de forma manual y validados por expertos. Además de estas placas, utilizamos un conjunto de imágenes digitalizadas en el Instituto de Ciencias Astronómicas, de la Tierra y del Espacio (ICATE; CONICET y Universidad Nacional de San Juan)¹². El etiquetado se realizó utilizando el software Label Studio.

¹² <https://icate.conicet.gov.ar/>

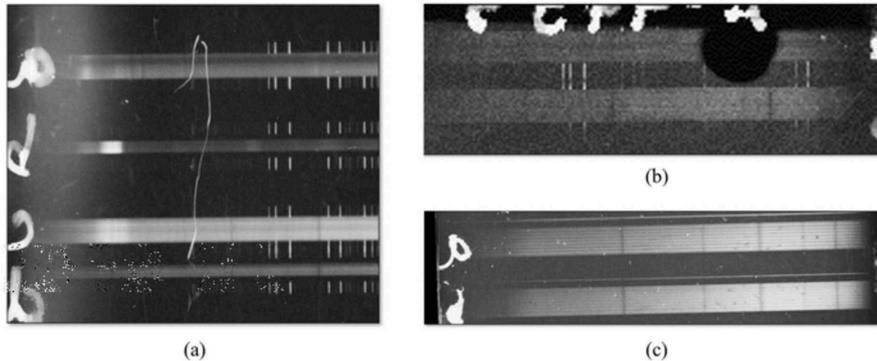


Figura 5. Ejemplos encontrados para la identificación automática de espectros. (a) y (b) suciedad o placas deterioradas. (c) espectros sin lámparas de comparación.

3.3. Modelo de detección

Como técnica de detección de los espectros entrenamos un modelo YOLO [10]. Este modelo, basado en Redes Neuronales Convolucionales, permite el entrenamiento en un dominio particular para detectar regiones de interés dentro de una imagen. Para realizar el entrenamiento redujimos las imágenes de sus tamaños originales a una resolución de 512×512 píxeles con una profundidad de color de 8 bits. Dividimos las imágenes en un 80 % para realizar el entrenamiento y 20 % para la evaluación. El modelo fue entrenado con tamaño de lotes de 64 y un total de 250 épocas. Utilizamos la técnica de *data augmentation* con los siguientes parámetros:

- Cambio de brillo $\pm 30\%$
- Ruido gaussiano del 10 % al 50 %
- Volteo completo de la imagen de forma vertical
- Rotación $\pm 3^\circ$
- Escalado $\pm 20\%$
- Mosaico

Los rangos operacionales de los parámetros fueron validados por un profesional de la astronomía para verificar que mantengan realismo. Luego del entrenamiento del modelo se logró obtener un *f-measure* cercano al 98 %, dando excelentes resultados de detección. La figura 6 muestra un ejemplo de detección en algunas placas del conjunto de imágenes para evaluación. Como puede apreciarse, en este caso la detección funcionó de manera aceptable. Los dos espectros no detectados (ver segunda placa de la figura) son espectros que no son válidos para analizar por estar corruptos, con lo cual el modelo se comportó como era deseado.

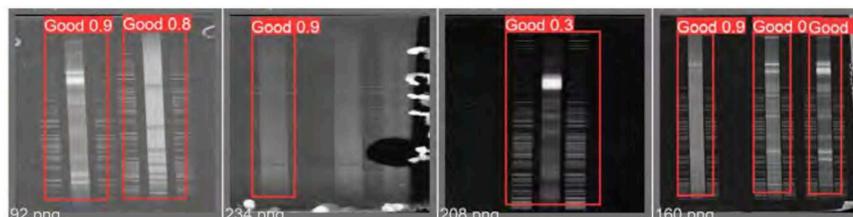


Figura 6. Validación del modelo entrenado en algunas imágenes del conjunto de Evaluación.

4. Conclusiones y trabajos futuros

En este trabajo presentamos el desarrollo de un software novedoso para asistir en la digitalización de placas espectroscópicas de vidrio. El software permite la digitalización de grandes cantidades de placas en menor tiempo, evitando errores y automatizando gran parte del proceso. Así, se posibilita efectivamente la conservación del patrimonio observacional, como así también el almacenamiento en formato adecuados de los espectros de ciencia para un futuro procesamiento. El software fue desarrollado utilizando Javascript para el frontend y Python para el backend, dando posibilidad de ejecutar el sistema desde cualquier navegador web.

Para agilizar el trabajo de las personas encargadas de la digitalización desarrollamos un detector automático de los espectros de dentro de cada placa digitalizada. Para esto, entrenamos un modelo basado en YOLO, creando una base de datos específica para este dominio. El modelo mostró excelentes resultados tanto con datos de validación como en nuevas placas digitalizadas.

Como trabajos futuros, proponemos:

- Validar el uso del software y del modelo de detección al digitalizar el total de placas existentes. Este trabajo ya se encuentra en proceso.
- Calibrar las lámparas de comparación es una tarea que el profesional debe hacer como primer paso en el procesamiento de los espectros de ciencia. Esta es una tarea manual y lenta que aplicarla a los miles de espectros existentes en la FCAyG sería inviable a corto plazo. Se está trabajando en la creación de modelos de aprendizaje automático para calibrar las lámparas de forma automática.
- Igual que para el inciso anterior, el procesamiento posterior del espectro es una tarea manual, donde se identifican las frecuencias de interés de un objeto particular. Como paso siguiente a la calibración de lámparas, se espera poder procesar la información existentes en las observaciones, obteniendo las diferentes curvas espectroscópicas para cada objeto.

Referencias

1. Calderón, J.H., Calderón, J.H., Bustos Fierro, I.H., Bustos Fierro, I.H., Melia, R., Willimoës, C., Willimoës, C., Giuppone, C., Giuppone,

- ne, C.: The Digital Archive of the Photographic Images of the Córdoba Observatory Plates Collections. *Ap&SS***290**(3), 345–351 (2004). <https://doi.org/10.1023/B:ASTR.0000032547.59015.7b>
2. Chollet, F., et al.: Deep learning with Python, vol. 361. Manning New York (2018)
 3. Cydale, L., Gamen, R., Aidelman, Y., Ronchetti, F., Quiroga, F., Rodriguez, J., López, M., Peralta, R., Meilán, N., Alessandroni, M.d.R., Colazzo, S., Pereyra, N.: Recuperación del trabajo observacional histórico (retroh). la facultad de ciencias astronómicas y geofísicas. unlp. <https://retroh.fcaglp.unlp.edu.ar/>, accedido el: 2/6/2022
 4. Heines, A.: Longterm Variability - First Results from Digitised Photographic Plates. In: Guenther, E., Stecklum, B., Klose, S. (eds.) *Optical and Infrared Spectroscopy of Circumstellar Matter*. Astronomical Society of the Pacific Conference Series, vol. 188, p. 171 (1999)
 5. Hudec, R., Kopel, F., Macsics, R., Hadwige, M., Heber, U., Cayé, W.: Classification of Variable Objects for Search for GRB Candidates on Bamberg Photographic Plates. *Acta Polytechnica* **53**(3), 27 (2013)
 6. Meilán, N., Collazo, S., Alessandroni, M.R., López Durso, M., Peralta, R.A., Aidelman, Y., Cidale, L.S., Gamen, R.: Proyecto de digitalización de placas espectrográficas del Observatorio de La Plata. *Boletín de la Asociación Argentina de Astronomía La Plata Argentina* **61B**, 251–253 (2020)
 7. Muminov, M., Yuldoshev, Q., Ehgamberdiev, S., Kahharov, B., Relke, H., Protsyuk, Y., Pakuliak, L., Andruk, V.: Astrometry of the η and χ Persei clusters based on the processing of digitized photographic plates. *Bulgarian Astronomical Journal* **26**, 3 (2017)
 8. Orellana, R.B., de Biasi, M.S., Bustos Fierro, I.H., Calderón, J.H.: A revisit to the region of Collinder 132 using Carte du Ciel and Astrographic Catalogue plates. *A&A***521**, A39 (2010). <https://doi.org/10.1051/0004-6361/200913741>
 9. Pence, W. D., Chiappetti, L., Page, C. G., Shaw, R. A., Stobie, E.: Definition of the flexible image transport system (fits), version 3.0. *Astronomy and Astrophysics* **524**, A42 (2010). <https://doi.org/10.1051/0004-6361/201015362>
 10. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016). <https://doi.org/10.1109/CVPR.2016.91>
 11. Rickman, H.: Unión astronómica internacional. b3 - safeguarding the information in photographic plates. <https://www.iau.org/static/publications/ib88.pdf>, accedido el: 2/6/2022
 12. Walborn, N.R., Gamen, R.C., Morrell, N.I., Barbá, R.H., Fernández Lajús, E., Angeloni, R.: Active Luminous Blue Variables in the Large Magellanic Cloud. *AJ***154**(1), 15 (2017). <https://doi.org/10.3847/1538-3881/aa6195>
 13. Wertz, M., Horns, D., Groote, D., Tuvikene, T., Czesla, S., Schmitt, J.H.M.M.: Hamburger Sternwarte plate archives: Historic long-term variability study of active galaxies based on digitized photographic plates. *Astronomische Nachrichten* **338**(1), 103–110 (2017). <https://doi.org/10.1002/asna.201613201>