

Desarrollo de una plataforma astroinformática para la administración y análisis inteligente de datos a gran escala

Detección y Clasificación Automática de Estructuras a Distintas Escalas Espaciales en Imágenes Astronómicas

Rodrigo Gregorio, Mauricio Solar, Diego Mardones,
Marcelo Mendoza, Karim Pichara, Ricardo Contreras, Victor Parada, Guillermo Cabrera,
Jorge Ibsen, Lars Nyman, Eduardo Vera, Paola Arellano, Paulina Troncoso.

Santiago, 29 de abril de 2014

Resumen

El estudio de las estructuras en es importante para la comunidad astronómica porque en ellas se pueden encontrar diversos objetos, los cuales requieren ser identificados y clasificados. Por este motivo, se propone el desarrollo de una aplicación automática que permita analizar imágenes y obtener una clasificación de estructuras astronómicas. La propuesta consiste en usar algoritmos de detección de estructuras (CLUMPFIND, Dendogramas, Wavelets) y luego aplicar algoritmos de clasificación. La finalidad es construir una herramienta que quede a disposición de la comunidad y cumpla con los requerimientos de un observatorio virtual.

Palabras Claves: ALMA, Imágenes Astronómicas, Detección, Clasificación

1. Resumen Ejecutivo

La detección, caracterización, y relaciones entre objetos representados por medio de una variedad de escalas espaciales y de propiedades químicas y físicas ofrece oportunidades únicas para el estudio del universo, en particular, usando los datos de ALMA. Por consiguiente, el desarrollo de herramientas computacionales que automaticen estas tareas es de enorme utilidad.

La propuesta consiste en desarrollar una herramienta para la detección automática y clasificación de estructuras a distintas escalas espaciales en imágenes astronómicas. Para ello primero se un filtro paso bajo provenientes de la transformada de wavelets para generar un subconjunto de imágenes con distintas escalas. La descomposición por medio de wavelets asegura que al unir el subconjunto de imágenes no se pierde información. La herramienta se aplica a imágenes en 2 dimensiones, dejando la posibilidad de aplicarla en una serie de planos en imágenes en 3 dimensiones. Luego, se aplican algoritmos de detección en cada subconjunto generando un catálogo para la región estudiada. El catálogo incluye una clasificación del tipo de objeto o estructura encontrada, y permite realizar estadísticas y buscar relaciones espaciales entre los objetos del catálogo. Esta herramienta se desarrolla en Python, debido a su facilidad para ser incluida en CASA (Common Astronomy Software Applications) y/o ChiVO (Chilean Virtual Observatory).

La herramienta creada es nueva e incluye el trabajo conjunto entre la astronomía y la ingeniería informática, y proporcionará una ayuda a la investigación científica y el desarrollo de nuevas aplicaciones en el área de la astro-informática.

Índice

1. Resumen Ejecutivo	2
2. Definición del Problema	4
3. Solución propuesta	4
3.1. Síntesis Hito 1	4
3.2. Transformada de Wavelet	5
3.2.1. Algoritmo A Trous	7
3.3. Algoritmo de Detección	9
3.3.1. Clumpfind	9
3.4. Árbol Jerárquico	10
3.4.1. Dendrogramas	11
3.5. Catálogo	11
3.6. Visualización	11
4. Trabajo futuro	13

2. Definición del Problema

La propuesta consiste en el desarrollo de un algoritmo para clasificar objetos analizando el tipo de dato entregado por el proyecto ALMA y la cantidad de objetos observados hasta el momento. En base a lo anterior se genera una propuesta que aborda temáticas que son de interés dentro de los involucrados en el proyecto de la creación de ChiVO.

El observatorio ALMA brindará datos en abundancia cuando comience su operación científica permanente, estos datos nos abrirán nuevas ventanas al estudio del universo. En parte debido a la sensibilidad a emisión de gas y polvo frío sin precedentes hasta la actualidad, y en particular debido a que los datos proveerán simultáneamente sensibilidad a:

- Estructuras espaciales a escalas desde 0.01 hasta 1000 segundos de arco simultáneamente.
- Una variedad de líneas espectrales simultáneas trazando estructuras de propiedades físicas y químicas muy diversas.

Esto permitirá estudiar la presencia de estructuras físicas sobre amplias escalas espaciales con diversas propiedades en cada tipo de objeto. Por lo anterior, la detección, caracterización, y relaciones entre objetos representados por una variedad de escalas espaciales y propiedades ofrece oportunidades únicas para el estudio del universo con los datos de ALMA. El desarrollo de herramientas computacionales que automaticen estas tareas es de enorme utilidad.

En el contexto de la creación de un Observatorio Virtual Chileno con los datos de ALMA, se hace necesario el desarrollo de herramientas de procesamiento y análisis de imágenes que permita hacer las labores de detección y clasificación automática de estructuras astronómicas en dichas imágenes.

3. Solución propuesta

3.1. Síntesis Hito 1

La solución propuesta consiste en aplicar distintos algoritmos de detección y clasificación. La figura 1 muestra el esquema de la propuesta.

La propuesta consiste en desarrollar un algoritmo que permita identificar y clasificar estructuras astronómicas a diferentes escalas, usando la transformada Wavelets. Se busca generar un subconjunto de imágenes a escalas distintas y luego aplicar algún algoritmo de detección (gaussclump, clumpfind, dendrogramas) para identificar objetos en cada imagen. Al unir todas las imágenes generadas se debiese obtener el conjunto total de estructuras que contiene la imagen que se está analizando (en formato FITS [3]).

A partir de las estructuras encontradas en el paso anterior se prosigue con la aplicación de un algoritmo de clasificación que permita reconocer a que tipo de objeto pertenecen las estructuras encontradas. El usuario (astrónomo) determinará que criterios se considerarán para clasificar estos objetos.

Como parte final, se espera generar una aplicación que se pueda utilizar de forma “standalone” o ser incluida dentro del software CASA(Common Astronomy Software Application) y/o estar dentro de

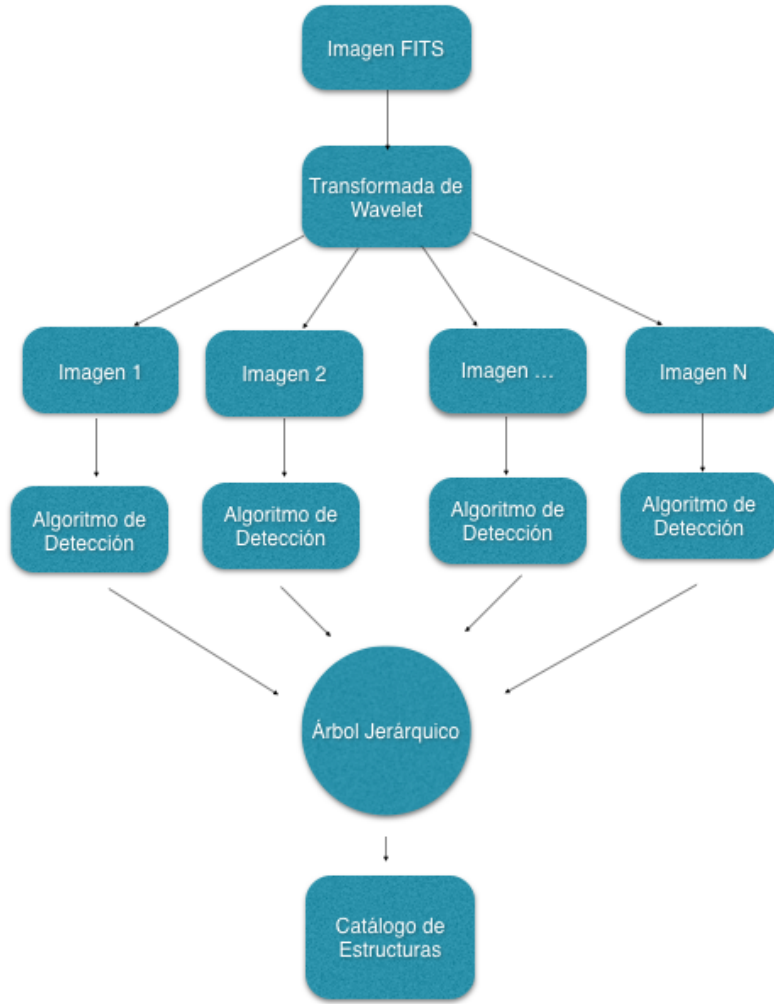


Figura 1: Proceso de creación de catálogo de estructuras a partir de una imagen FITS

ChiVO. De esta forma la herramienta generada estará al alcance de la comunidad y además debe cumplir con los requerimientos de un Observatorio Virtual.

3.2. Transformada de Wavelet

En el libro de Starck y Murtagh [5] se hace una revisión de las distintas aplicaciones de las Wavelets en astronomía. Además, en el apéndice se entrega una descripción del algoritmo para la implementación de una transformada de Wavelet discreta conocida como A Trous. En [2, 1] se dan ejemplos de aplicaciones de las Wavelets en astronomía.

Las funciones wavelet son la base para poder aplicar su transformada, la cual permite descomponer señales jerárquicamente y su posterior reconstrucción, además de extraer ciertos tipos de información de

ellas. Las wavelet son herramientas para la descomposición de señales, tales como imágenes en una jerarquía de resoluciones crecientes de tal manera que al considerar niveles de mayor resolución, se obtengan más y más detalles de la imagen.

Una cualidad muy importante de las funciones wavelet es poder analizar una señal a diversas escalas. En el análisis wavelet, la escala juega un papel muy importante, ya que los diversos algoritmos procesan los datos en diferentes escalas y resoluciones espaciales. Si se observa una señal a una escala pequeña (poca resolución) se apreciarán únicamente aquellas entidades de mayores dimensiones. De igual manera, si se observa la misma señal a una escala grande (alta resolución) se podrán distinguir pequeños elementos o partes de la señal. Por ello, analizar una señal $f(t)$ consiste en descomponer en una serie de versiones escaladas y trasladadas con el objetivo de representarla como la superposición de un conjunto de funciones base o wavelet escaladas y trasladadas.

En la práctica se aplica una función wavelet prototipo llamada “wavelet madre” a partir de la cual se deriva toda una familia de versiones trasladadas y escaladas. El análisis de una señal en función del tiempo $f(t)$ se realiza desde dos puntos de vista: el análisis temporal, mediante una versión de la función madre contraída y de alta frecuencia; mientras que el análisis de frecuencias se desarrolla con una versión dilatada y de baja frecuencia. Como la citada señal original $f(t)$ puede ser representada en términos de una expansión wavelet, **“las operaciones sobre los datos pueden ser realizadas empleando solo los correspondientes coeficientes wavelet”**.

La forma más sencilla y frecuente de poner en práctica las propiedades que las funciones wavelet poseen para el estudio imágenes consiste en aplicar convoluciones sobre ellas mediante filtros cuyos coeficientes son derivados de las funciones wavelet. Ese conjunto de filtros, tanto de descomposición como de reconstrucción o síntesis, se denomina banco de filtros.

Una idea básica para procesos de filtrado es que la distribución de energía en el dominio de la frecuencia identifica a una estructura. Por tanto, si el espectro de frecuencia es descompuesto en un número suficiente de subbandas, la energía de diferentes estructuras serán desiguales. Aprovechando esta cualidad, se han diseñado varios tipos de bancos de filtros y entre ellos cabe citar los filtros separables de Laws (1980), los filtros circulares y en cuña (Coggins y Jain, 1985).

Un filtro digital es una secuencia de valores que se emplea para destacar o suavizar ciertos aspectos en una señal, sea de una o dos dimensiones. Se aplica sobre una señal mediante una convolución produciendo otra señal de salida diferente. El filtro es desplazado sobre la señal calculando un producto interno (producto punto) entre los coeficientes del filtro y aquellos puntos de la señal sobre los que se encuentra el filtro. La representación digital de un filtro es conocida como respuesta de impulso y aquellos filtros que tienen un número finito de coeficientes son llamados filtros de respuesta de impulso finita. Los filtros con un número infinito de coeficientes se denominan filtros de impulso infinito.

Los filtros digitales pueden ser simétricos o asimétricos. Los filtros simétricos y especialmente los que tienen una forma con pico en el centro, tienen una serie de ventajas: preservan la localización de las transiciones agudas en las señales y facilitan el tratamiento de sus bordes. Los filtros simétricos son a veces llamados de fase lineal, ya que si no lo son, su desviación es evaluada por la magnitud de desviación de su fase desde una función lineal. Además, existen dos tipos de filtros simétricos, los simétricos respecto del valor central del filtro (tienen dimensión impar) y por tanto sus coeficientes cumplen la relación $h(k) = h(-k)$; y los simétricos en mitad del filtro (dimensión par), que cumplen la relación

$h(k) = h(-k - 1)$. Los filtros asimétricos cumplen que $h(k) = -h(-k)$ o bien $h(k) = -h(-k - 1)$.

Uno de los inconvenientes de la Transformada Wavelet Discreta (TWD) radica en que no es invariante a las traslaciones, es decir, una imagen inicial y otra en la que se haya realizado una pequeña traslación, presentarán diferentes coeficientes wavelet en la transformación. Este contratiempo es importante en aplicaciones como detección de bordes, determinación de patrones espaciales y reconocimiento de imágenes en general.

Se puede evitar este efecto aplicando una transformada wavelet redundante o no dividida (Undecimated Wavelet Transform), de dos formas posibles: (1) no dividiendo la imagen de entrada, tan solo operando sobre ella mediante los correspondientes bancos de filtros, de tal forma que el subconjunto de imágenes resultantes de la transformada wavelet tengan las mismas dimensiones que la imagen inicial; o bien (2) aplicando sobre ella y sobre sus versiones desplazadas la TWD ordinaria (que implica submuestreo), consiguiendo una redundancia de información que es equivalente a no dividirla.

3.2.1. Algoritmo A TrouS

Fue desarrollado por Holschneider, Kronland-Martinet, Morlet y Tchamitchian, en 1989. Consiste en una descomposición basada en la transformación wavelet discreta en la que no se produce submuestreo de las imágenes, sino que éstas presentan siempre la misma resolución. Se realiza una convolución con los filtros básicos paso bajo H y paso alto G que son expandidos insertando un número apropiado de ceros entre los coeficientes. De forma práctica se descompone la imagen convolucionándola mediante un filtro paso bajo bidimensional, obteniendo de esta manera una imagen de aproximaciones, mientras que los coeficientes de detalles resultan de la diferencia entre dos imágenes consecutivas filtradas con el citado filtro paso bajo. De forma inversa se procedería para la reconstrucción.

El filtro bidimensional de paso bajo generalmente consiste en un filtro spline bi-cúbico, asociado a la función scaling, si bien se pueden aplicar otros filtros cuyos coeficientes sean los correspondientes a los aplicados en la descomposición wavelet discreta. En la figura 2 se muestra la rutina de cálculos para obtener las imágenes de aproximaciones y detalles a diferentes niveles, con un filtro genérico de paso bajo de 5x5.

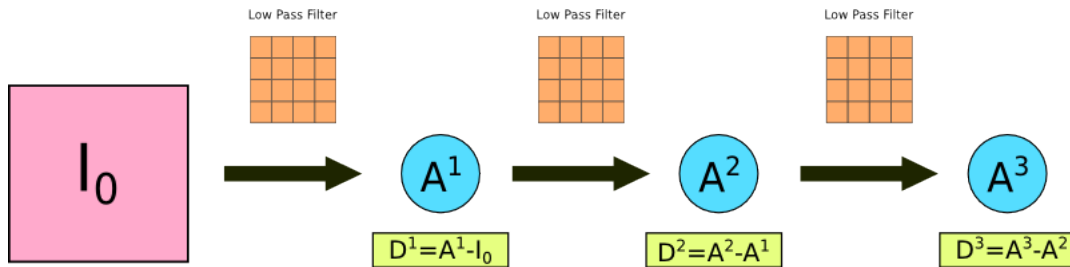


Figura 2: Algoritmo “à trous” aplicado sobre una imagen de partida I_0 e imágenes resultantes

El algoritmo “à trous” emplea de un filtro B3-spline discreto de tamaño 5x5 en el que los coeficientes quedan definidos según la tabla 1, con todos sus coeficientes multiplicados por $1/256$.

Tabla 1: filtro B3-spline

1	4	6	4	1
4	16	24	16	4
6	24	36	24	6
4	16	24	16	4
1	4	6	4	1

La distancia entre los valores del filtro en cada nivel se incrementa con un factor de 2 (figura 3).

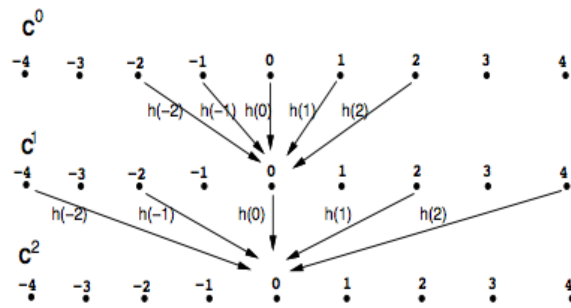


Figura 3: Paso del nivel c_0 , c_1 , c_2

Una prueba se realiza sobre la región NGC 6334 (ver figura 4). En la figura 5 se ven 8 niveles de descomposición.

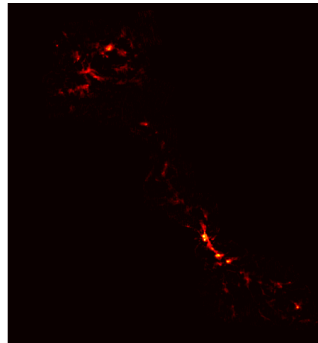


Figura 4: Región NGC 6334

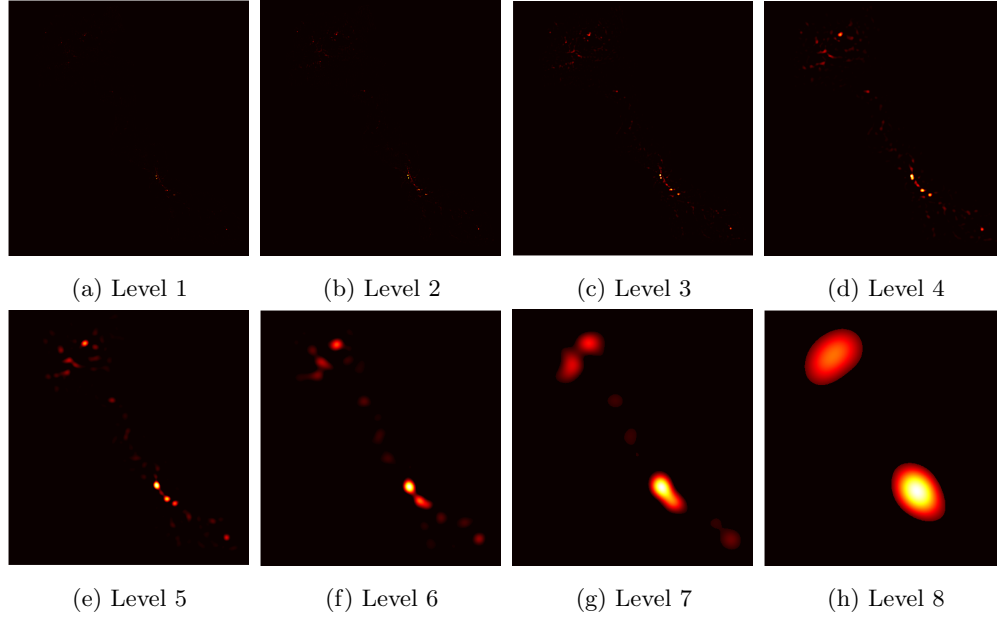


Figura 5: Imágenes 8 niveles obtenidos con la transformada de wavelet

3.3. Algoritmo de Detección

3.3.1. Clumpfind

Clumpfind [6] es un algoritmo automático para el análisis de la estructura en una línea espectral de cubo de datos. El algoritmo funciona mediante la construcción de contornos, siendo múltiplos del ruido rms (root mean square) de las observaciones, a continuación, se busca los picos de emisión como nuevos clumps. A cada emisión se le asigna clump, se sigue así hasta la menor intensidad. Fue propuesto por Williams et al. (1994), esta basado en como el ojo analizaría los mapas: se contornea el conjunto de datos, buscando picos, y se continua con los niveles de contornos más bajos de forma secuencial.

Los problemas fundamentales son: como establecer los niveles de contornos para los datos y como manejar el caso de cuando dos o más clumps se mezclan.

Los contornos deben ser espaciados, $T = \Delta T, 2\Delta T, 3\Delta T, \dots$, ya que el ruido se agrega linealmente en cada nivel. Si ΔT es muy pequeño, el mapa de contornos aparecerá lleno de estructuras dificultando la diferenciación entre características reales y picos de ruido. Por otro lado, cuando ΔT es muy grande el mapa de contornos carecerá de contraste y características sutiles se perderán. En las pruebas realizadas por Williams et al. (1994) establecieron un valor adecuado para $\Delta T = 2T_{rms}$, donde $2T_{rms}$ es el ruido rms en la imagen.

El algoritmo define como un clump a la colección de píxeles en el cual su más alto contorno esta aislado de cualquier otro clump, es decir, no están conectados. Los clump deben ser aislados en un mismo nivel de contorno, sin embargo, se mezclarán en niveles menores (ver figura 6). Se crea un vector que contiene en cada item la colección de píxeles del clump i , cada nuevo clump es un nuevo item en el vector. Los resultados sobre la región NGC 6334 se ven en la figura 7

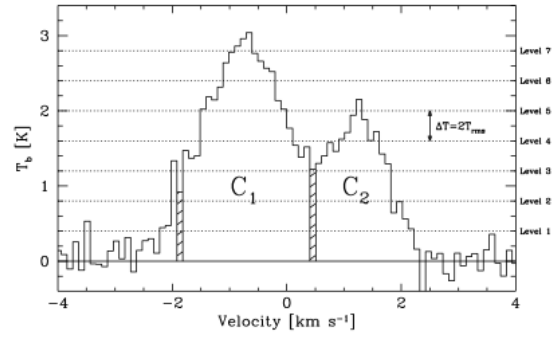


Figura 6: Espectro con niveles de contorno

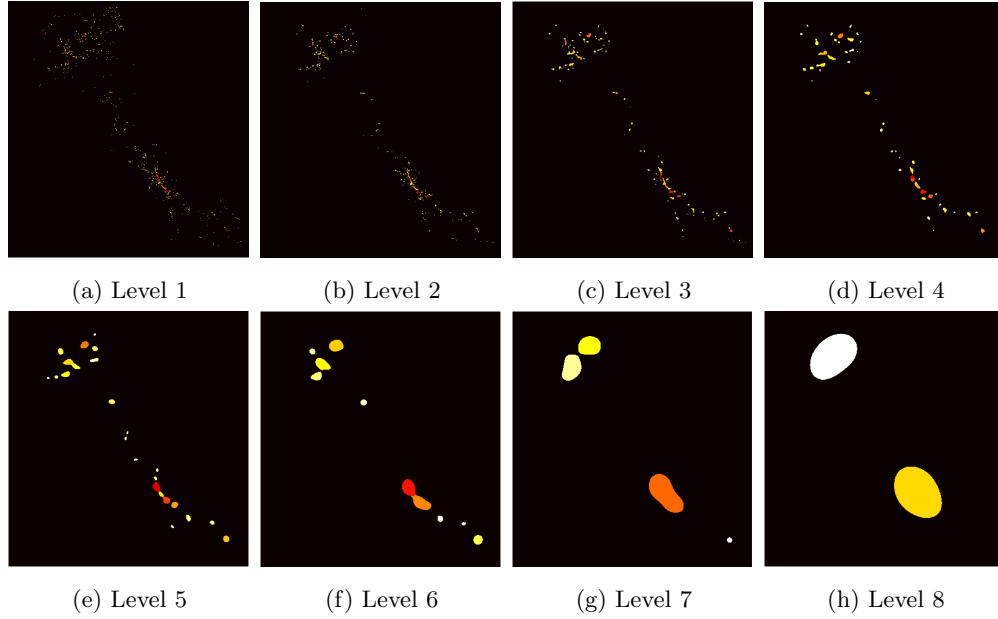


Figura 7: Imágenes de la detección en 8 niveles

Para estos casos se descarta el último nivel el cual se interpreta como ruido.

3.4. Árbol Jerárquico

Para ver las relaciones entre las distintas estructuras encontradas en cada nivel se hace uso de estructuras jerárquicas tipo árbol.

3.4.1. Dendrogramas

Un dendrograma [4] es un tipo de representación gráfica en forma de árbol que organiza los datos en subcategorías y que se van dividiendo en otras hasta llegar a un nivel de detalle deseado. Este tipo de representación permite apreciar claramente las relaciones de agrupación entre los datos e incluso entre grupos de ellos aunque no las relaciones de similaridad o cercanía entre categorías. Observando las sucesivas subdivisiones podemos hacernos una idea sobre los criterios de agrupación de los mismos, la distancia entre los datos según las relaciones establecidas, etc. También podríamos referirnos al dendrograma como la ilustración de las agrupaciones derivadas de la aplicación de un algoritmo de clustering jerárquico.

A partir del algoritmo clumpfind implementado se obtienen la posición de los picos más altos en cada clump y también la posición de su centroide. Con los datos anteriores se puede construir un dendrograma por cada nivel. Para encontrar relaciones entre los niveles se necesitan estructuras jerárquicas que consideren solo la ubicación de los picos o centroide y no su valor, de esta forma se relacionan las estructuras entre niveles. Se determina que clumps están dentro de otro entre cada nivel, esto es similar a los dendrogramas pero solo se considera la pertenencia para establecer las hojas.

3.5. Catálogo

Por cada imagen FITS que se desea examinar se genera un catálogo de estructuras encontradas. Este catálogo contiene los datos de posición y pico mas alto centroide (ver Tabla 2).

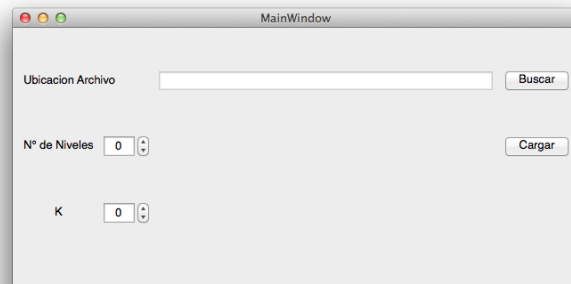
Tabla 2: Catálogo generado con imagen en nivel 6

ID	area[px]	x_{peak}	y_{peak}	$valor_{peak}$	$x_{centroide}$	$y_{centroide}$	$valor_{centroide}$
n6_0	4367	400	744	0.564	400	746	0.562
n6_1	4646	330	797	0.268	324	808	0.240
n6_2	3411	1114	381	0.240	1114	380	0.239
n6_3	3221	1025	307	0.126	1020	311	0.122
n6_4	1816	139	1094	0.098	139	1094	0.098
n6_5	1774	961	285	0.078	960	283	0.077
n6_6	697	1085	261	0.055	1085	260	0.055
n6_7	745	829	518	0.053	829	518	0.053
n6_8	737	245	904	0.052	245	904	0.052
n6_9	273	220	1022	0.041	220	1022	0.041

3.6. Visualización

Para tener un mejor manejo con el algoritmo se implementa una interfaz de usuario complementaria a la propuesta. La cual consiste en desplegar distintas ventanas para ingresar los datos y configurar los parámetros del algoritmo.

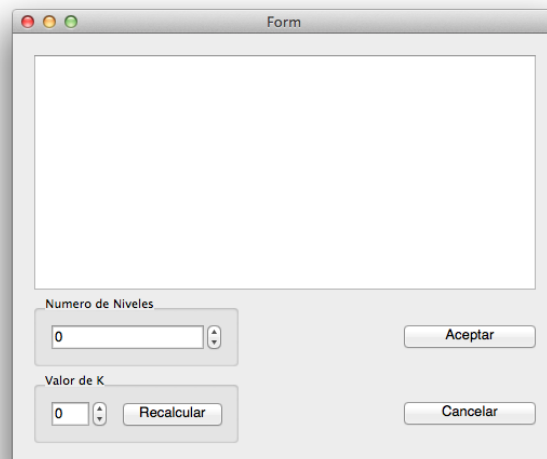
En la vista de ingreso de datos (ver figura 8) se entrega la dirección de la imagen FITS, además se selecciona la cantidad de niveles que se quiere evaluar (estos niveles están restringidos por el tamaño de la imagen y el filtro paso bajo de la transformada wavelet).



The 'MainWindow' dialog box contains three input fields: 'Ubicacion Archivo' (a text box), 'Nº de Niveles' (a spin box set to 0), and 'K' (a spin box set to 0). To the right of the first field is a 'Buscar' button, and to the right of the second field is a 'Cargar' button.

Figura 8: Ventana de ingreso de imagen FITS

La vista previa (figura 9) es una ventana que permite ver los niveles resultantes permitiendo tener una vista anterior del proceso del algoritmo. Esto ayuda al usuario a determinar una cantidad adecuada de niveles según sus especificaciones.



The 'Form' dialog box features a large empty rectangular area at the top for a preview. Below this, there are two input sections: 'Numero de Niveles' with a spin box set to 0 and an 'Aceptar' button, and 'Valor de K' with a spin box set to 0 and a 'Recalcular' button. A 'Cancelar' button is located at the bottom right of the dialog.

Figura 9: Ventana de vista previa de los niveles de la transformada de wavelet

4. Trabajo futuro

El trabajo que queda pendiente puede resumirse en:

1. Desarrollar una forma de visualización del árbol jerárquico entre los niveles wavelet. Así mismo, desplegar los dendrogramas por nivel.
2. Desarrollar las ventanas para los dendrogramas y el árbol jerárquico, y la ventana para el catálogo.
3. Integrar la interfaz de usuario y el algoritmo propuesto. Luego, ver su inclusión en ChiVO (Chilean Virtual Observatory).
4. Realizar pruebas sobre un conjunto de imágenes mas amplio.

Referencias

- [1] J. Alves, M. Lombardi, and C. J. Lada. The mass function of dense molecular cores and the origin of the IMF. *Astronomy & Astrophysics*, 462:L17–L21, January 2007.
- [2] Stéphane Jaffard, Yves Meyer, and Robert D. Ryan. *12. Wavelets and Astronomy*, chapter 12, pages 187–201. Society for Industrial and Applied Mathematics, 2001.
- [3] R. J. Hanisch, A. Farris, E. W. Greisen, W. D. Pence, B. M. Schlesinger, P. J. Teuben, R. W. Thompson, and A. Warnock III. Definition of the flexible image transport system (fits). *Astronomy & Astrophysics*, 376(1):359–380, 2001.
- [4] E. W. Rosolowsky, J. E. Pineda, J. Kauffmann, and A. A. Goodman. Structural Analysis of Molecular Clouds: Dendrograms. *The Astrophysical Journal*, 679:1338–1351, June 2008.
- [5] J.-L. Starck and F. Murtagh. *Astronomical Image and Data Analysis*. Springer, 2nd edition, 2006.
- [6] J. P. Williams, E. J. de Geus, and L. Blitz. Determining structure in molecular clouds. *The Astrophysical Journal*, 428:693–712, June 1994.