RECONOCIMIENTO DE PATRONES PARA LA IDENTIFICACIÓN DE CLASE Y FAMILIA DE PLANTAS A PARTIR DE SUS CARACTERES

Ricardo Rodrigo Juárez Hernández[1], José Sergio Ruiz Castilla[2], Jair Cervantes Canales[3], Farid García Lamont[4]

1 guillermus@prodigy.net.mx Maestrante UAEM
2 jsergioruizc@gmail.com Profesor de Tiempo Completo UAEM
3 chazarra17@gmail.com Profesor de Tiempo Completo UAEM
4 fglamont@yahoo.com.mx Profesor de Tiempo Completo UAEM

Resumen

La clasificación de plantas en base a sus características se le llama taxonomía o proceso taxonómico, ccomprende tres actividades principales que son: clasificación, nomenclatura e identificación. En la nomenclatura se le asigna un nombre científico a las diferentes especies acorde al Código Internacional de Nomenclatura Botánica. En el Instituto Botánico de la Universidad Nacional Autónoma de México (UNAM) existe un "museo de plantas", el más completo de México. Las plantas almacenadas primero son "curadas" por una "autoridad" en el ramo. Cuándo se recolecta una planta repetidamente se almacena tomando en cuenta el lugar de origen y la fecha de recolección, para estudiar el comportamiento de la especie en el tiempo y espacio. El proceso anterior es manual coordinado por el Dr. José Luís Villaseñor Ríos. El problema existente se origina porque se tienen miles de muestras pendientes por identificar y debido al alto grado de complejidad y tiempo requerido. Por lo que se plantea como objetivo desarrollar un algoritmo para la identificación de plantas del Instituto Botánico de la UNAM acorde a los métodos usados por el curador mencionado para las cerca de 21 776 especies identificadas en México. Este sistema deberá considerar el catálogo de 150 características que pueden poseer o no cada familia. Se aplicará la metodología marcando las características que tienen una planta y mediante reconocimiento de patrones se determinará el grado de similitud que tiene con otras especies.

Palabras clave

Clase, familia, algoritmo, reconocimiento de patrones, especies, plantas.

Abstract

Plants classification is based from their characteristics is called taxonomy or taxonomic process, comprehends three mayor activities, they are: classification, nomenclature and identification. At the nomenclature a scientific name is assigned to the different species according to the International Code of Botanic Nomenclature. In the Instituto Botánico de la UNAM exist a “Plants Museum”, the most complete from México. The stored plants go through the following process: first they are prepared (curadas) for the field expert. When a plant is recollected continuously is stored saving the information from the origin place and the recollection date, in order study to the behavior of the species in that time and geographic zone. The previous process is manual coordinated by the Dr. José Luis Villaseñor Rios. The actual problem is because they are thousands of pending species to identify due the high degree of complexity and time required. The objective of this research is to develop an algorithm of plants identification from the Instituto Botánico de la UNAM according to the used methods for the field expert for the 21 776 species identified so far in México. This system should consider a catalog of 150 characteristics that each plant family might have. The applied methodology will be marking the characteristics from a plant and by patterns will be determinate the similarity degree from other existing species.

Key Words

family, algorithm, patterns recognition, species, plants.

I Introducción

La identificación taxonómica se define como el proceso de nombrar o catalogar un espécimen dentro de un sistema de clasificación previa. Comprende tres actividades principales que son: clasificación, nomenclatura e identificación (Villaseñor, 1993). La clasificación es el proceso de asignar un espécimen o grupo de especímenes dentro de una categoría taxonómica, o sea un taxón. En la nomenclatura se le asigna un nombre a las diferentes categorías acorde al Código Internacional de Nomenclatura Botánica

(Villaseñor, 1998). Por último, en la identificación se determina dónde debe estar ubicado un espécimen dentro de este sistema de clasificación internacional botánico.

México es uno de los países más ricos en diversidad. Su flora incluye cerca de 21 777 especies diferentes y en su territorio se centra una gran diversidad de grupos taxonómicos. Por ejemplo dentro de éstas familias se encuentra la *Asteracea*, que consiste en un número de más de 1 400 géneros y cerca de 23 000 especies (K, 1994). En México esta planta se clasifica en 310 a 387 géneros y de 240 a 300 especies, dependiendo de la autoridad taxonómica consultada (Villaseñor, 1993).

En el "museo de plantas" del Instituto de la UNAM es el más completo de México. Las plantas almacenadas primero son "curadas" que consiste en clasificarla correctamente por un experto para su almacenamiento. Cualquier planta repetida se almacena guardando la fecha y lugar donde se recolectó. El proceso anterior es manual coordinado por la autoridad en el ramo que es el Dr. José Luís Villaseñor Ríos.

Debido a la modalidad del proceso, se tienen miles de plantas por identificar y clasificar debido a la escasez de personal, tiempo y recursos. Anteriormente este Instituto desarrolló una herramienta para ayudar en esta tarea de clasificación llamada GENCOMEX: a *Computerized Key to Identify the Genera of Asteraceae Of México* (Villaseñor, 1998), desarrollada en 1998 en el lenguaje de programación Pascal y sólo corre en Sistema Operativo Windows de 32 bits, por lo que lleva un rezago de ésta herramienta de 15 años al menos.

El uso de la computadora en la identificación biológica apoya en dos tareas, la primera, como auxiliar en la elaboración automatizada de claves para identificación, y como un medio en sí, para llevar a cabo la identificación (Munguía-Romero, 1992).

Para identificar una planta taxonómicamente es necesario conocer a la familia que pertenece (Munguía-Romero, 1992). Para algunas plantas es fácil diagnosticar su familia, pero para la mayoría de las plantas clasificarlas en familias es una tarea difícil ya que requiere tiempo y conocimientos en el ramo.

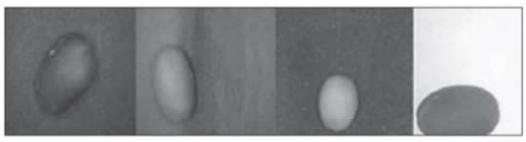
Con el algoritmo propuesto se logran identificar las plantas a través de sus características con porcentajes de precisión de hasta el 95%.

Se revisaron algunos trabajos de investigación con el fin de conocer el estado del arte y se encontraron diversos, de los cuales se hace referencia a algunos de ellos. En el trabajo de Sistema de selección electrónico de café excelso basado en el color mediante procesamiento de imágenes de investigación de Ruge, Pinzón y Moreno describen una aplicación del procesamiento de imágenes digitales mediante la técnica de umbralización multinivel para la selección del café de una manera automática. Es necesario hacer uso de recursos tecnológicos que den solución a problemas reales, como los procesos de selección de café para dar un producto de mayor calidad al mercado.

El objetivo de este trabajo es proponer una solución relativamente económica y más ajustada para fincas cafetaleras de baja y mediana producción, haciendo uso del procesamiento de imágenes.

El procedimiento que llevaron a cabo es el siguiente: se tomó un color específico y uniforme de café. Se van dosificando los granos de café uno a uno, estos granos pasan a una tolva donde se les toma una foto con una resolución de 320x240 pixeles. Posteriormente se seleccionó el fondo de pantalla más adecuado tomando en cuenta el brillo, tono y color del café seleccionado previamente dando como resultado el color blanco véase la Fig. 1.

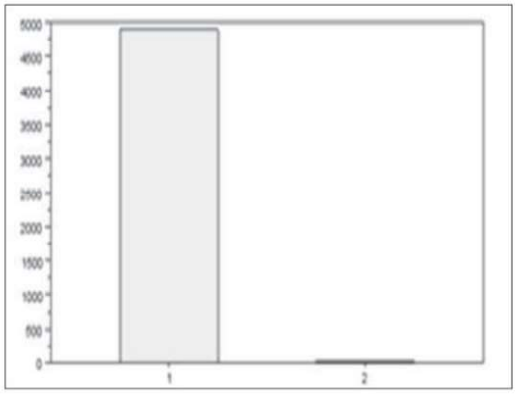
Fig. 1 Pruebas de fondo para toma de imagen a procesar (Ilber, 2012)



Con el fin de ahorrar tiempo computacional se detecta solo una parte para saber si hay un grano que analizar o no. Posteriormente se estableció un umbral con un valor de 130 que es el rango de las tonalidades que puede tener un grano de café, la cámara tiene una iluminación constante y está ubicada a 7 cm del grano. Esta foto es procesada y transformada a escala de grises para poder realizar su histograma mediante IMLAB. En la segmentación multinivel se asignan un rango de valores que pueden tener los granos de café entrantes para poder analizar un rango amplio de granos de café.

Los resultados que se obtuvieron fueron en granos buenos de café, histogramas muy consistentes hacia un color definido, tal como se muestra en la Fig. 2, la máquina selectora de granos de café por medio de un dispositivo mecánico separa en diferentes contenedores los granos que considera buenos y de los malos.

Fig. 2 Histograma de la imagen de un grano bueno de café (Ilber, 2012)



Para validar los resultados probaron con 100 granos de café, hicieron la prueba con el sistema clasificador, acto seguido el experto realizó también la selección manual de los granos dando resultados favorables acorde a la clasificación del experto.

En las conclusiones hacen hincapié en una iluminación uniforme en la imagen, porque se establecen los umbrales a utilizar y los cambios de luz afectan notablemente los resultados de forma negativa (Ilber, 2012).

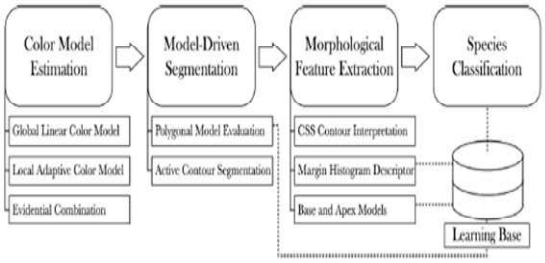
En base a este artículo, la aportación es el utilizar de manera concreta y eficiente la técnica de umbralizado multinivel, y estos resultados son comparados con un experto en la materia como retroalimentación del sistema.

En otro trabajo de Entendiendo hojas en imágenes naturales-un método basado en modelo para identificación de especies de árboles por Cerutti, Taugne, Mille, Vacant y Coquin se enfocan en hacer una herramienta educativa, que se apoye en criterios de geometría de alto nivel inspirado el uso de los botánicos, que hacen una posible interpretación semántica, para clasificar una hoja dentro de una lista de especies.

Su objetivo fue construir un sistema (llamado “Folia”) para el análisis de la forma de una hoja que sea procesada en ambiente natural.

Para la investigación utilizaron el siguiente procedimiento: primero hacen un análisis de segmentación y modelos de contornos activos. Posteriormente se centran en trabajar una sola especie, en este caso hojas palmatilobadas, las fotos son tomadas en ambiente natural por una cámara móvil y la hoja está centrada en la foto y orientada verticalmente. Acto seguido proponen patrones preestablecidos de forma de hoja basada en la abstracción de la interpretación del botánico. Después estiman el modelo del color de la hoja para determinar que puede ser parte de la hoja y que no basado en un modelo polinomial proporcionado por la forma de la hoja. Después obtienen las características discriminatorias de la especie por ejemplo su margen, base, ápex, etc. Y finalmente interpretan en contorno final usando el modelo de espacio escala-curvatura. La Fig. 3 muestra este procedimiento de manera global.

Fig. 3 Diagrama del proceso general de identificación de hojas (Cerutti, 2012)



Para probar el sistema obtuvieron imágenes y datos de *Pl@ntLeaves Database*, donde utilizaron dos tercios de los datos como conjunto de entrenamiento y el tercio restantes como prueba, en esta se incluyen fotos con fondo blanco o en su medio ambiente.

Los resultados obtenidos de este conjunto de datos son los siguientes de manera general: en imágenes escaneadas con un 82.4% de efectividad, pseudo-escaneadas con 83.5 %, fotografías con 71.6 %, obteniéndose un promedio de 79.2 % de efectividad.

Concluyen que el sistema propuesto es una base sólida de donde comenzar para reconocimiento de hojas de árboles en su ambiente natural. Por otra parte la introducción

de descriptores inspirados en botánica de alto nivel es alcanzable en el sector educativo (Cerutti, 2012).

En base a este artículo, la aportación es no dejar todo el proceso de clasificación al dispositivo móvil razones de petición-espera y por otra parte utilizar todas las herramientas y conocimientos disponibles por parte del experto del ramo (en este caso el botánico) para realizar una identificación más eficiente, tales como características morfológicas y color de las hojas.

II Materiales y métodos

II.1. Materiales

Con el fin de comprobar la eficiencia en cualquier computadora sin ninguna prestación especial se utilizó un *procesador AMD* E-300 1.3 *Gigahercios (Ghz)* con 2 *Gigabytes (Gbytes)* de *Random Access Memory (RAM)* para la experimentación. El software utilizado fue *Java Runtime Enviroment 7 (JRE 7)*.

II.2. Métodos

II.2.1. Procedimiento de clasificación.

En base a los procedimientos establecidos por el Dr. José Luis Villaseñor Ríos se determina como se identifican los grupos de Clases y Familias en México. La primera gran división es entre 2 clases que son Liliopsida (4500 especies) y Magnoliopsida (17500 especies). Dentro de la clase Liliopsida se encuentran 2 grandes familias que son la Orchidaceae que posee cerca de 1500 especies y la Poaceae con 1000 especies aproximadamente. Por la parte de la clase de Magnaliopsida que es el conjunto más grande se encuentran 2 grandes familias que son la Asteraceae con cerca de 9000 especies y la Fabaceae con 2000 especies aproximadamente. El resto de las familias de cada clase, género y especie de cada una se identificaran por sus características.

II.2.2. Disposición de la información fuente

La información proporcionada por el Instituto de Botánica de la UNAM está dispuesta primero por una lista de 21776 especies clasificadas por clase, familia y género. Ver Tabla 1.

Tabla 1. Plantas clasificadas.

	Clase	Familia	Género	Especie
1	Liliopsida	Alismataceae	Echinodorus	andrieuxii
2	Liliopsida	Alismataceae	Echinodorus	berteroi
3	Liliopsida	Alismataceae	Echinodorus	cordifolius
4	Liliopsida	Alismataceae	Echinodorus	grandiflorus
5	Liliopsida	Alismataceae	Echinodorus	nymphaeifolius
6	Liliopsida	Alismataceae	Echinodorus	paniculatus
7	Liliopsida	Alismataceae	Echinodorus	tenellus
8	Liliopsida	Alismataceae	Echinodorus	virgatus
9	Liliopsida	Alismataceae	Helanthium	Bolivianum
…	…	…	…	…
21776	Magnoliopsida	Zygophyllaceae	Viscainoa	Geniculata

Por otro lado se tienen 150 características posibles que posee cada familia. No se cuenta aún con las características de todas las especies aunque se está trabajando para caracterizar a nivel de especie. Como las características están numeradas, entonces es posible leer como un vector de caracteres de 150 elementos. Ver la Tabla 2.

Tabla 2. Características de las familias.

	Características
1	Plantas leñosas (árboles o arbustos)

2	Plantas herbáceas (anuales o perennes, incluyendo sufrútices)
3	Bejucos o plantas escandentes
4	Plantas acuáticas o subacuáticas
5	Plantas epífitas
6	Plantas parásitas o saprófitas
7	Plantas con jugo lechoso (látex)
8	Plantas con jugo acuoso (no lechoso)
9	Plantas aromáticas o resinosas (en corteza, ramas u hojas)
10	Plantas con zarcillos
…	…
150	Vegetación acuática

II.2.3. Caracterización de las hojas

De igual forma y siguiendo con el procedimiento establecido por el Dr. Villaseñor, se usó la convención de 1's si posee una característica y 0's si hay ausencia de dicha característica dentro de las 150 posibles características, resultando un vector de 150 valores con 1's y 0´s. La Tabla 3 muestra las características de cada familia.

Tabla 3. Familias con sus características.

	Clase	Familia	001	002	003	004	005	…	150
1	Liliopsida	Agavaceae	1	1	0	0	0	…	0
2	Liliopsida	Alismataceae	0	1	0	1	0	…	1
3	Liliopsida	Alliaceae	0	1	0	0	0	…	1
4	Liliopsida	Aloeaceae (I)	1	1	0	0	0	…	0
5	Liliopsida	Alstroemeriaceae	0	1	1	0	0	…	0
6	Liliopsida	Amaryllidaceae	0	1	0	0	0	…	1

II.2.4. Diseño del algoritmo

Se crea el archivo con los vectores de los datos de la fila que corresponde a cada familia. Debido a que un gestor de base de datos puede alargar el tiempo de respuesta entre la aplicación (CONNOLLY, 2005), se decidió cargar la información directamente a la memoria *RAM* para optimizar la velocidad de procesamiento. El pseudo-código para realizar el experimento en las familias elegidas se muestra en el Código 1

```
Inicio
        Carga el programa principal
        Crea el objeto tipo Plant
        Crea una lista tipo iterador de plantas
        Lee las plantas del archivo
        Crea todos los objetos tipo Plant
        Carga los objetos creados a la lista
        Por cada objeto Plant de la lista:
             Compara una a una cada característica con
             el objeto planta a comparar.
        Imprime resultados
Fin
```

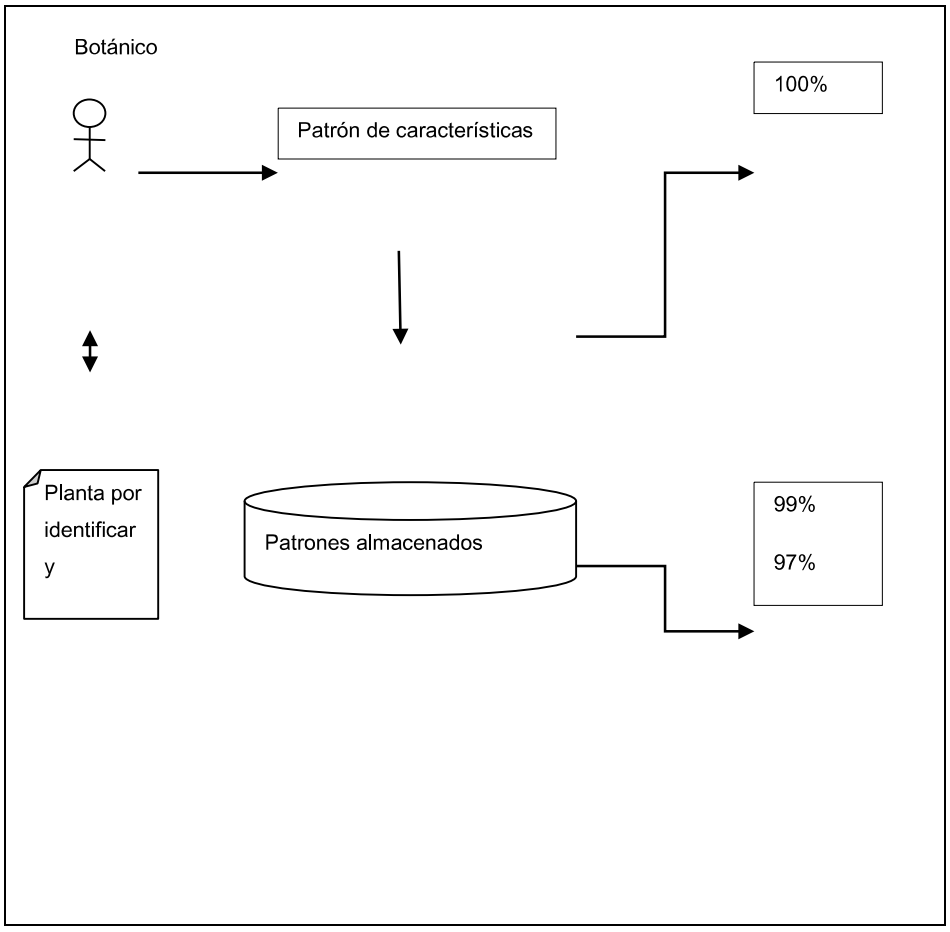
El programa principal llamado TaxonV2 tiene un método que crea y cargar la clase *DBPlants*. La clase *DBPlants* que se forma de una lista (colección) de objetos de 0..N de la clase *Plant* es creada en ese momento sin ningún objeto. Por otra parte la colección *DBPlants* posee los métodos para agregar, eliminar, devolver una planta y devolver el número de objetos propios de la lista para un óptimo uso de la clase. Posteriormente el programa principal manda a leer el archivo con las plantas y sus respectivos vectores renglón de sus características. Acto seguido el programa instancia una clase *Plant* por cada renglón leído del archivo y se agrega a la colección *DBPlants*. La clase *Plant* tiene el atributo privado de características donde se representa el vector fila que posee cada planta, también tiene los métodos de acceso y mutadores (*getters* y *setters*) propios de cada clase por principio de Encapsulación. Terminado de leer todo el archivo la lista *DBPlants* se regresa al programa principal con todos los objetos ya cargados. Posteriormente se compara las características de la planta a buscar encapsulada en

objeto tipo *Plant* con la lista directamente *DBPlants.* Finalmente imprime los resultados de las comparaciones.

II.2.5. Estrategia de proceso experimental.

Se ha optado por usar un algoritmo para integrar un patrón de una familia considerando 150 características de 0's y 1's en una sola cadena. El botánico interesado en identificar y clasificar una planta debe determinar las características con valor de 1's en las posiciones correspondientes. El nuevo patrón es procesado contra los patrones existentes almacenados. El resultado obtenido puede ser del 100% si coincide en su totalidad con algún patrón almacenado. Si faltan características por definir los resultados serán porcentajes más bajos de acuerdo a la similitud con otras familias. Ver figura 4.

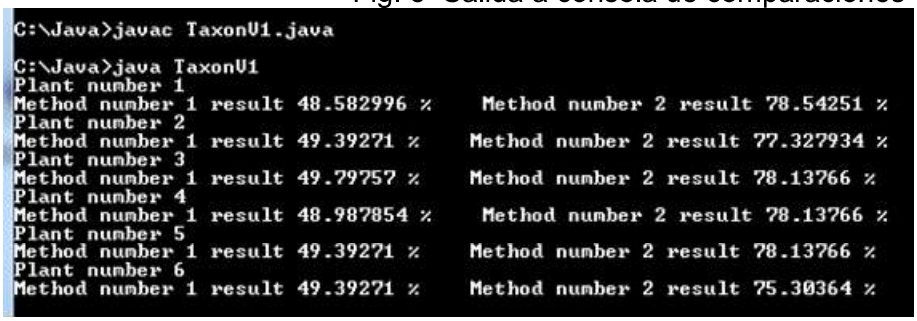
Fig. 4 Esquema de la estrategia del proceso experimental



III. Resultados

Se realizaron ejecuciones aumentando características para conocer el comportamiento, cuando el botánico no introduce o bien no cuenta o no conoce todas las características de la planta. Se muestra una salida de las ejecuciones en la Figura 5.

Fig. 5 Salida a consola de comparaciones



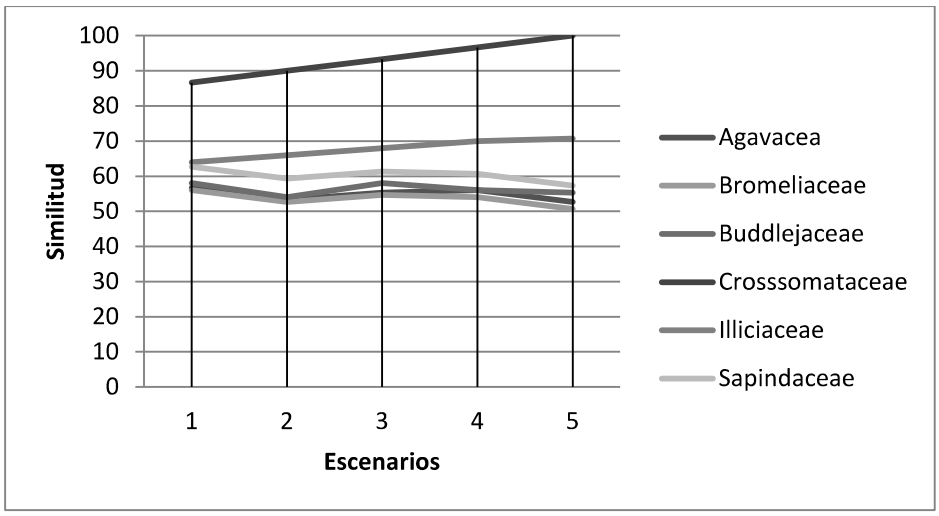
Se muestra con 6 familias con 150 características diferentes cada una. Ver Tabla 4.

Tabla 4. Datos de ejecuciones del algoritmo.

Clase	Familia	1	2	3	4	5
Liliopsida	Agavaceae	56.66	53.33	55.33	56	52.66
Liliopsida	Bromeliaceae	56	52.66	54.66	54	50.66
Magnoliopsida	Buddlejaceae	58	54	58	56	55.33
Magnoliopsida	Crossosomataceae	86.66	90	93.33	96.66	100
Magnoliopsida	Illiciaceae	64	66	68	70	70.66
Magnoliopsida	Sapindaceae	62.66	59.33	61.33	60.66	57.33

La figura 6 nos permite observar el comportamiento de resultados de similitud con los patrones: Para la familia correcta converge hacia el valor 100%, mientras que para las familias que no corresponde disminuye alejándose y dejándose de ser familias candidatas. Finalmente el curador podrá verificar si la planta corresponde a la familia que el algoritmo nos ha arrojado.

Fig. 6. Comportamiento de reconocimiento de patrones.



Por la parte de tiempos se realizó la prueba de acceso a RAM, comparando cada planta con una colección de 5000 plantas que es un promedio de búsqueda dentro de cada conjunto de clases y familias se obtuvieron los siguientes resultados: se tardó 375 milisegundos.

IV. Conclusión

Una vez llevada a cabo la experimentación se ha probado el algoritmo obteniendo resultados que permiten plenamente determinar un porcentaje más alto de la familia correcta. Sobre todo cuando no se cuentan con todas las características de la planta.

Cuando no se consideran todas las características la precisión no es del 100% pero si es superior a los demás patrones, por lo que el algoritmo cumple el objetivo de ayudar al Botánico a identificar a que familia pertenece una determinada planta.

V. Discusión y trabajo futuro.

Debido a que actualmente en la UNAM el sistema anterior solo clasifica clase y familia, se desarrollará como trabajo futuro un sistema usando un algoritmo que servirá de base para identificar y clasificar clase, familia, género y especie de alguna planta, con el fin de obtener la lista completa de 21 777 especies hasta el momento identificadas y clasificadas por el Dr. Villaseñor.

La intención del presente trabajo es dejar un sistema base eficaz y eficiente para una completa identificación y clasificación, y por otra parte, perfeccionar la herramienta en base al tiempo, uso y retroalimentación por parte del usuario final.

VI. Referencias

CERUTTI Guillaume, L. T. (2013). Understanding leaves in natural images- A model-based approach for three species identification. *ELSEVIER Computer Vision and Image Understanding,* 117, 1482-1501.

JOLY Alexis, H. G.-F. (2013). Iteractive plan identification based on social image data. *ELSEVIER Ecological Informatics,* 413, 1-13.

MUNGUÍA-ROMERO, J. L. (1992). La computadora en la identificación botánica. *La era digital Ciencia y Desarrollo*.

PERTOR Ilaria, T. K. (2012).). Identificador: A web based tool for fisual plant disease identification, a proof of concept with a case study on strawberry. *ELSEVIER Computer and Electronics in Agriculture,* 84, 144-154.

RUGE RUGE ILBER ADONAYT, P. A. (2012). Sistema de selección electrónico de café excelso basado en el color mediante procesamiento de imágenes. *Tecnura*, 84-93.

VILLASEÑOR, J. L. (1993). La familia Asteraceae en México. *Revista de la Sociedad Mexicana de Historia Natural*, 117-124.

VILLASEÑOR, M. M. (1998). GENCOMEX: a computarized key to identify the genera of asteraceae of México. *Asociación de Biólogos de la Computación AC*.

CONNOLLY Thomas, Begg Carolyn (2005), Database Systems, A practical approach to Desing, Implementation, and Managment. 4th Edition, Ed Addison_Wesley an imprinting of Pearson Education, 29-30.