Búsqueda por Similitud de Marcas de Ganado Vacuno

Lucrecia Michel[1], Rita Romani[1], Andrés Pascal[1]

[1] Dpto. de Sistemas de Información, Universidad Tecnológica Nacional, Entre Ríos, Argentina,

{lucrecia.a.michel, ritaromani90, andrespascal22}@gmail.com

Abstract. *Similarity searching is an important field of study in the present days. An example of its application is the process of seeking cattle brands that is required during the registration of a trademark. In the present article, a method of similarity searching of cattle brands is presented. This method uses a variation of the Hausdorff distance that improve the precision of the results.*

Resumen. *Las búsquedas por similitud constituyen un campo de estudio de gran importancia en la actualidad. Un ejemplo de su aplicación es la búsqueda de marcas de ganado vacuno requerida durante el proceso de registro de una marca. En el presente trabajo se presenta un método de búsqueda por similitud de marcas de ganado que utiliza una variación de la distancia de Hausdorff para mejorar la precisión de sus resultados.*

1. Introducción

Las búsquedas sobre bases de datos tradicionales se basan en el concepto de búsqueda exacta: la base de datos es dividida en registros, teniendo cada registro campos completamente comparables; una consulta a la base retorna todos aquellos registros cuyos campos coincidan por igualdad con la consulta, o a lo sumo verifique su relación mediante operadores relacionales como mayor o menor. En la actualidad, muchas aplicaciones tienen como necesidad buscar en grandes bases de datos objetos que sean similares a uno dado. En este tipo de búsqueda no tienen sentido los operadores relaciones y se requieren otras formas de comparación. Estas búsquedas reciben el nombre de *Búsquedas por Similitud* y tienen diversos campos de aplicación [Navarro, Baeza-Yates y and Marroquin 2001].

En este artículo nos enfocamos en la búsqueda por similitud de imágenes que representan marcas de ganados, necesidad detectada en el Registro de Marcas y Señales de la Subsecretaría de Agricultura, Ganadería y Pesca del Ministerio de Agroindustrias de la provincia de Buenos Aires [Ministerio de Agroindustrias], El registro de marcas se encuentra regulado por el [Decreto Ley Nacional 22939 SENASA 1983], que establece en su artículo 3: *"no se admitirá el registro de diseños de marcas iguales, o que pudieran confundirse entre sí, dentro del ámbito territorial de una misma provincia o Territorio*

Nacional. Se comprenden en esta disposición las que presenten un diseño idéntico o semejante, y aquellas en las que uno de los diseños, al superponerse a otro, lo cubriera en todas sus partes". Esta regla genera la necesidad de contar con algún mecanismo de búsqueda por similitud de dichas marcas, ya que realizarlo manualmente es inseguro e ineficiente considerando que existen alrededor de 55000 marcas registradas.

El problema de búsqueda de imágenes puede abstraerse en un universo de objetos U y una función de distancia *d* que modela la similitud entre los objetos del universo. El par (U, d) se denomina espacio métrico. La base de datos será un subconjunto finito $X \subseteq U$. Esta problemática usualmente implica un preprocesamiento de las imágenes que se almacenarán la base de datos y de las imágenes de consulta para obtener algún tipo de estructura que represente sus características.

Existen antecedentes en la investigación en búsquedas por similitud de imágenes de ganado vacuno. Una de ellas se basa en el cálculo de histogramas de las pendientes de la imagen sobre los cuales se aplica el coeficiente de Pearson para su comparación [Sampallo, Duarte, Vázquez y González Thomas 2003]. El problema del uso de las pendientes es que es altamente sensible a la rotación, inclusive ante pequeñas variaciones. Además los resultados mostrados no son concluyentes. Otra línea de investigación propone su resolución mediante la generación de histogramas que representan las geometrías de las marcas basados en las distancias entre pares de puntos aleatorios de la imagen, Posteriormente se utiliza una distancia de la familia de las distancias de Minkowski para comparar los vectores [Torres y Rodríguez García 2014]. Este enfoque es muy interesante pero necesita mayor verificación para establecer si es representativo del nivel de similitud. En este artículo proponemos un método alternativo: el cálculo de la similitud a partir de la función de distancia de Hausdorff [Moreno, Koppal and de Muinck 2013][Arguello 2008] y de una variación de la misma, para aumentar la precisión de los resultados obtenidos.

Este artículo está organizado de la siguiente manera: en la Sección 2 se exponen los fundamentos teóricos del trabajo; en la Sección 3 se describe nuestra propuesta; la Sección 4 muestra resultados experimentales y finalmente, en la Sección 5 se exponen las conclusiones y futuras líneas de trabajo.

2. Trabajo Relacionado

2.1. Espacios Métricos

El problema de búsquedas por similitud entre objetos puede abstraerse como un universo de objetos *U* y una función de distancia *d* que modela la similitud entre los mismos.

Un Espacio Métrico se define como un par *(U, d)* donde *U* es el universo de objetos válidos del espacio y $d: U \times U \rightarrow R+$ es una función de distancia que mide el grado de similitud (disimilitud, en realidad) entre los elementos de *U*. Esta función *d* cumple con las propiedades características de una función métrica: positividad $(\forall x, y \in U, d(x, y) \geq 0)$, reflexividad $(\forall x \in U, d(x, x) = 0)$, simetría $(\forall x, y \in U, d(x, y) = d(y, x))$ y desigualdad

triangular *(∀x, y, z ∈ U, d(x, y) ≤ d(x, z) + d(z, y))*. La base de datos es cualquier subconjunto finito $X \subseteq U$.

Una consulta típica por similitud en espacios métricos es la búsqueda por rango, que se denota $(q, r)_d$, donde q es un elemento del universo U, al que se denomina *query*, y r un radio de tolerancia. Una búsqueda por rango consiste en recuperar todos los objetos de la base de datos que estén a lo sumo a distancia r de q, es decir: $(q, r)_d = \{x \in X / d(q, x) \leq r\}$.

La importancia del uso de espacios métricos como modelo, es que permite el uso de índices que permiten que la búsqueda se realice con mayor eficiencia que *O(n)* evaluaciones de distancias, que es el costo de comparación de la consulta con todos los elementos de la base de datos.

2.2. Distancia de Hausdorff

La función de distancia que utilizamos en nuestro trabajo es la Función de Distancia Hausdorff (Ecuación 1) [Moreno, Koppal and Muinck 2013][Arguello 2008], que se aplica a dos conjuntos finitos de elementos (puntos, en nuestro caso) y se define como:

$$DH(A,B) = \max\{h(A,B), h(B,A)\}$$

Ecuación 1. Distancia de Hausdorff

donde h(A,B) es la distancia de Hausdorff Directa, y se calcula mediante la Ecuación 2.

$$h(A,B) = \max_{a \in A}\{\min_{b \in B}\{d(a,b)\}\}$$

Ecuación 2. Distancia de Hausdorff Directa

$A = \{a_1, a_2, a_3, \dots, a_k\}$, $B = \{b_1, b_2, b_3, \dots, b_p\}$ son conjuntos; *d(a, b)* es una distancia métrica tal como la Distancia Euclidiana, y max(x) y min(x) son funciones que calculan los valores máximos y mínimos respectivamente [Arguello 2008].

Es importante resaltar que la distancia h(A, B) es distinta a la distancia h(B, A). Comúnmente el conjunto A representa un patrón que se desea encontrar dentro del objeto representado por B, por este motivo a la distancia h(A, B) se denomina distancia *directa* y h(B, A) es llamada distancia *inversa* de Hausdorff. La Distancia de Hausdorff se puede utilizar para comparar imágenes en blanco y negro a través de los conjuntos de puntos que representan las posiciones de sus pixeles negros. En este artículo hacemos uso de esta función y una variación de la misma que mejora la calidad de los resultados.

3. Método de Búsqueda por Similitud de Marcas de Ganado

Las imágenes correspondientes a las marcas se almacenan en la base de datos junto a una representación de las mismas a través de un conjunto de puntos. Para ello, cada imagen se

procesa previamente, como veremos más adelante. Cuando se ingresa una imagen de consulta, ésta se transforma también en un conjunto de puntos y luego se realiza la búsqueda. Ya que las funciones de distancia utilizadas son métricas, se pueden utilizar índices tales como el FHQT, FQA, LAESA, etc. [Chavez, Navarro, Baeza-Yates, y Marroquin 2001], para acelerar la búsqueda.

3.1. Procesamiento de las Imágenes.

La transformación de las imágenes de las marcas que se almacenan en la base de datos, se realiza mediante los siguientes pasos:

a. Binarización: las imágenes suelen estar en formato color o escala de grises, por lo cual en un primer paso se las transforma a blanco y negro, mediante la fijación de un umbral.
b. Esqueletización: este proceso se realiza con el fin de convertir las líneas a un solo pixel de ancho y reducir significativamente la cantidad de puntos que la componen [Carranza 2006][Eberly 2001], ya que en este caso el ancho de las líneas es irrelevante para la evaluación de la forma.
c. Obtención del Minimum Bounding Rectangle (MBR): consiste en determinar el área rectangular mínima que contiene todos los puntos de la imagen [Eberly 2015].
d. Redimensión de la imagen a una medida estandarizada de 50 pixels de ancho, para que todos los elementos de la base de datos se encuentren en la misma resolución.

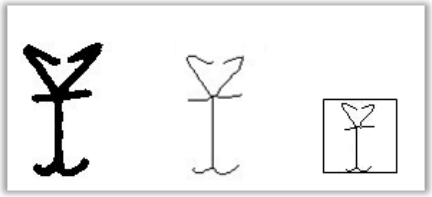
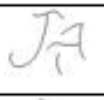


Figura 1. binarización, esqueletización, MBR y Redimensión.

e. Rotaciones y Espejados de la imagen (Tabla 1): para que las búsquedas sean robustas ante rotaciones, de decidió incorporar a la base de datos copias de cada imagen (imágenes asociadas) con distintos grados de rotación, que mantienen su relación con la imagen principal a través de una clave foránea. También se realizó el espejado de la imagen y las rotaciones correspondiente.
f. Obtención y almacenamiento del conjunto de puntos negros de cada imagen asociada.

Tabla 1. Marcas de Ganado Rotadas y Espejadas

Marcas de Ganado	0°	45°	90°	135°	180°	225°	270°	315°
Rotadas								
Espejadas								

3.2. Algoritmos y Funciones de Búsqueda por Similitud

Como mencionamos anteriormente, la función métrica utilizada fue, en principio, la Distancia de Hausdorff para medir el grado de similitud entre dos imágenes representadas a través de conjuntos de puntos.

Una falencia que presenta esta función es que cualquier punto alejado de los demás puntos de la imagen puede aumentar considerablemente el valor de la función, interpretándose como que las imágenes son distintas, a pesar de coincidir en todos los demás puntos. En la Figura 2, se muestra este problema. Las marcas son altamente similares pero la segunda posee un punto alejado que modifica significativamente el resultado de la distancia de Hausdorff.

Figura 2. Marcas similares, una de ellas con un punto lejano agregado.

Para solucionar este problema y obtener mayor precisión en los resultados, en la Función de Distancia de Hausdorff Directa se reemplazó el cálculo del máximo por el cálculo del promedio de los mínimos. Entonces la Distancia Promedio de Hausdorff se define mediante la siguiente expresión:

$$h(A,B) = \underset{a \in A}{\text{prom}} \min_{b \in B}\{d(a,b)\}$$

Ecuación 3. Distancia Promedio de Hausdorff

Además de realizar esta adaptación, se incorporó un algoritmo similar a una búsqueda binaria para encontrar posibles traslaciones de la imagen. El algoritmo realiza una primera comparación entre la consulta y el elemento de la base de datos en su posición original y dos comparaciones más con la imagen trasladada hacia izquierda y derecha un cierto valor de desplazamiento. Luego si alguna de las comparaciones de los extremos es menor que la central, se divide el intervalo entre ese extremo y el centro por dos y se realiza una nueva comparación trasladando la imagen a ese punto. Esta subdivisión se realiza como máximo tres veces. En algunos casos este procedimiento produce mejoras importantes en los resultados de la comparación de los objetos.

4. Resultados Experimentales

Para verificar el comportamiento del método de búsqueda propuesto, se realizaron diferentes pruebas sobre un conjunto de marcas provistas por Departamento de Registro Ganadero dependiente de la Dirección Provincial de Carnes del Ministerio de Agroindustria de la Provincia de Buenos Aires.

La base de datos utilizada se conformó a partir de 202 marcas originales que mediante los procesos de rotación y espejado se extendieron a 3.232 marcas asociadas. Se desarrolló un lote de marcas de consulta de 50 elementos. Estas marcas se obtuvieron de la siguiente manera: 40 tomadas de la base de datos y modificadas agregando y quitando puntos y líneas y realizando desplazamientos y rotaciones. Las otras 10 se diseñaron desde cero como marcas nuevas.

Para cada elemento de consulta se calcularon los tres vecinos más cercanos utilizando la distancia de Hausdorff (DH), la distancia Promedio de Hausdorff (DPH) y la distancia Promedio de Hausdorff con el agregado del algoritmo de búsqueda binaria de traslaciones (DPHT). En la Tabla 2 se muestran los resultados de la aplicación de las tres funciones al lote de consultas, tomando en cuenta sólo el vecino más cercano, y en la Tabla 3, considerando los aciertos dentro de los tres vecinos más cercanos.

Tabla 2. Cantidad y porcentaje de aciertos en el primer lugar

Función	Aciertos	%
DH	20	40.00
DPH	38	76.00
DPHT	38	76.00

Tabla 3. Cantidad y porcentaje de aciertos en los tres primeros lugares

Función	Aciertos	%
DH	21	42.00
DPH	38	76.00
DPHT	40	80.00

Si bien el porcentaje de acierto utilizando la distancia de Hausdorff es bajo, los resultados obtenidos mediante los otros dos métodos son suficientemente buenos como para continuar su análisis y mejora para la aplicación real al problema de la búsqueda de marcas de ganado. El algoritmo de traslación no produce una mejora significativa y aumenta considerablemente el costo de la búsqueda, por lo cual la distancia Promedio de Hausdorff es la distancia más conveniente de las tres. En la práctica, una alternativa para incrementar el porcentaje de aciertos es simplemente aumentar la cantidad de vecinos más cercanos recuperados, ya que en cualquier caso el personal del Registro de Marcas es el que decide si la marca buscada es suficientemente similar a otra ya registrada o no, y este método se puede utilizar como herramienta para reducir significativamente las marcas a considerar para tomar esa decisión.

5. Conclusiones y Trabajo Futuro

En este artículo se propone un método completo que utiliza una variación de una función de distancia de Hausdorff en conjunción con algoritmos y procedimientos que permiten la búsqueda por similitud de marcas de ganado con suficiente precisión como para ser utilizados, con algunas mejoras, en una aplicación real. El método utiliza una distancia métrica que permite el uso de índices para que la búsqueda sea más eficiente. Además se describen experimentos realizados para verificar su precisión y se exponen los resultados.

Actualmente estamos trabajando en la mejora del método para incrementar su precisión y disminuir el costo del cálculo de la función de distancia. Por otro lado, tenemos planificadas las siguientes tareas para un futuro próximo:

a. Comparación de nuestro método contra las técnicas nombradas de histogramas de pendientes e histogramas de distancias de pares de puntos, para la misma base de datos y lote de consultas.
b. Realización de experimentos con una base de datos de tamaño similar a la real (55000 marcas) y con la incorporación de un índice métrico, para evaluar eficiencia.
c. Diseño de un método alternativo de comparación de imágenes basado en la descomposición de las líneas de la imagen en trazos más simples.

6. Referencias

E. Chavez, G. Navarro, R. Baeza-Yates, and J.L. Marroquin (2001). Searching in metric spaces. ACM Computing Surveys, 33(3): 273–321

Ministerio de Agroindustrias de la Provincia de Buenos Aires. http://www.maa.gba.gov.ar

Decreto Ley Nacional 22939 SENASA (1983). Servicio Nacional de Sanidad y Calidad Agroalimentaria. Título I - De las marcas y señales en general. (artículos 1 al 4)

G. Sánchez Torres y M. E. Rodríguez García (2014). Medida de similaridad entre imágenes de marcas de ganado mediante distribuciones de forma, Revista Ingenierías Universidad de Medellín, vol. 13, No. 25 ISSN 1692 - 3324 - julio-diciembre de 2014/248 p. Medellín, Colombia.

G. Sampallo, D. Duarte, D. Vázquez y A. González Thomas (2003). Reconocimiento de Marcas de Ganado. Facultad Regional Resistencia, Universidad Tecnológica Nacional. CACIC.

Rodrigo Moreno, Sandeep Koppal, Ebo de Muinck (2013). Robust estimation of distance between sets of points. Center for Medical Image Science and Visualization (CMIV), Linköping University, Sweden. Department of Medical and Health Sciences (IMH), Linköping University, Swede.

Henry ARGUELLO (2008). Comparación de Huellas Dactilares Usando la Distancia Hausdorff, Facultad de Ingenierías Físico Mecánicas, Universidad Industrial de

Santander Bucaramanga, Santander 68001000, Colombia, SISTEMAS, CIBERNÉTICA E INFORMÁTICA VOLUMEN 5 - NÚMERO 2 - ISSN: 1690-8627

Carranza Athó, Fredy (2006). Tópicos Especiales en Procesamiento Gráfico. Escuela Académico Profesional de Informática, Universidad Nacional de Trujillo. Trujillo, Peru.

David Eberly (2015). Minimum-Area Rectangle Containing a Set of Points. Geometric Tools, LLC.

David Eberly (2001). Skeletonization of 2D Binary Images. Geometric Tools, LLC. June 7.