



Búsqueda por Similitud en Bases de Datos Multimedia

Benjamin Bustos

*CIW – PRISMA Research Group
 Departamento de Ciencias de la Computación
 Universidad de Chile*


Motivación

- Buscador de imágenes en la Web
 - Problema: Encontrar imágenes relevantes a una cierta descripción (consulta)
 - Por ejemplo: una huella digital
 - Encontrar cualquier huella digital
 - Yahoo! (imágenes)
 - Búsqueda utiliza metadatos
 - Consulta = "huella digital"

09/11/2009 Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia 2

Motivación

- Identificar la huella digital
 - Problema mucho más difícil
- Búsqueda basada en contenido
 - Utilizar imagen de la huella para realizar la búsqueda

Consulta = 

09/11/2009 Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia 3

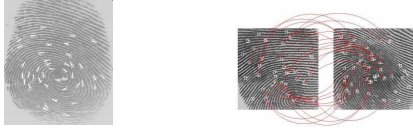
Motivación

- Dificultades:
 - ¿Pueden ser dos lecturas de la huella digital (incluso de la misma persona) exactamente iguales?
 - R: es improbable que suceda eso
 - 1 bit distinto => imágenes diferentes
 - Búsqueda exacta no es útil
 - Sin embargo: huellas deberían ser parecidas
 - **Búsqueda por similitud**

09/11/2009 Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia 4

Motivación

- Más dificultades
 - ¿Cómo se pueden comparar dos huellas digitales?
 - R: fijarse en características importantes de la huella
 - **Modelo de similitud (dependiente de la aplicación)**



09/11/2009 Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia 5

Motivación

- Aún más dificultades:
 - ¿Cómo se pueden buscar las huellas digitales más parecidas?
 - Solución ingenua: búsqueda secuencial
 - Lector estándar USB: 40.000 huellas/segundo
 - Sistema con 1 millón de usuarios: 25 segundos
 - **Algoritmos y estructuras de datos eficientes que apoyen la búsqueda por similitud**

09/11/2009 Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia 6

Agenda

- Tipos de datos multimedia
- Búsqueda por similitud
- Modelos de similitud para datos multimedia
- Temas de investigación

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

7

Tipos de datos multimedia

- Texto
 - A menudo simplemente: *strings* (secuencia de caracteres)
 - Documento
 - Párrafo
 - Capítulo de libro
 - Página Web
 - Requerimientos
 - Tolerancia a faltas de ortografía (fallas de OCR)
 - Búsqueda de documentos similares

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

8

Tipos de datos multimedia

- Imagen (estática)
 - Fuente: cámara digital, escáner
 - Muchos formatos distintos
 - Muchas formas de compresión (*lossless*, *lossy*)
- Conceptualmente: una matriz de puntos (píxeles)

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

9

Tipos de datos multimedia

- Modelos 3D
 - Modelos tridimensionales de objetos geométricos
 - Estructura: típicamente mallas de triángulos
 - Con o sin orientación (interior, exterior)
 - Cerrados, abiertos
 - Propiedades adicionales: textura, color, etc.
 - Requerimientos de espacio variables
 - Número de polígonos utilizados
 - Resolución deseada

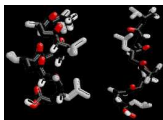
09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

10

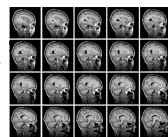
Aplicaciones

- BD multimedia: muchas aplicaciones prácticas



Biología molecular

Medicina



Geografía



Industria manufacturera

Y muchas otras...

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

11

Agenda

- Tipos de datos multimedia
- Búsqueda por similitud
- Modelos de similitud para datos multimedia
- Temas de investigación

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

12

Conceptos básicos

- Concepto de **similitud** es inherentemente subjetivo
- Modelos para definir objetivamente “similitud”
 - *Modelo general:*
 - Similitud: parte concordante entre objetos
 - Características concordantes conducen a X% similitud
 - Problema de optimización: “¿Cuánto cuesta transformar un objeto en otro?”

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

13

Conceptos básicos

- Modelos para definir objetivamente “similitud”
 - *Similitud basada en distancias:*
 - Función de distancia mide **disimilitud** entre objetos
 - A mayor distancia, más disímiles los objetos
 - Se puede formalizar matemáticamente
 - **Espacios métricos**
 - **Espacios vectoriales**

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

14

Espacios métricos

- Espacio vectorial: caso particular de espacio métrico
- \mathbb{R}^d : d -tuplas de números reales (*vectores*)

$$x \in \mathbb{R}^d \Rightarrow x = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_d \end{bmatrix}, x_i \in \mathbb{R}$$

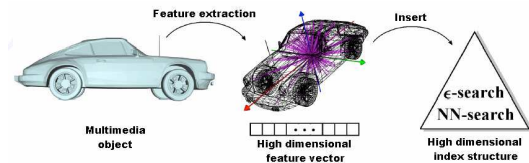
09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

15

Espacios vectoriales

- Método de vectores característicos



- Búsqueda: buscar puntos cercanos en un espacio vectorial

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

16

Consultas por similitud

- Sea \mathbb{U} el conjunto de datos
 - **Consulta por rango**
 - Objeto de consulta: $q \in \mathbb{X}$
 - Radio de tolerancia: $r \in \mathbb{R}^+$
- $$(q, r) = \{u \in \mathbb{U}, \delta(u, q) \leq r\}$$
- **Bola de consulta:** subespacio definido por q y r

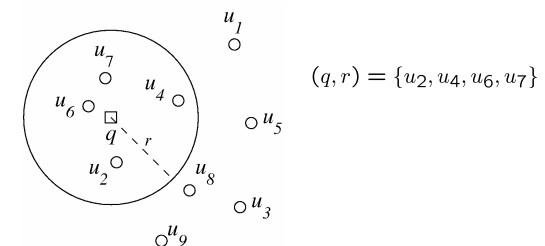
09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

17

Consultas por similitud

- Ejemplo de consulta por rango en (\mathbb{R}^2, L_2)



09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

18

Consultas por similitud

- Problema de la consulta por rango: ¿Qué valor debe tener el radio de tolerancia?



Muy pequeño: no retorna nada

Muy grande: retorna demasiado

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

19

Consultas por similitud

- Consulta por vecinos más cercanos (k -NN)
 - Objeto de consulta: $q \in \mathbb{X}$
 - Número de vecinos: $k \in \mathbb{N}$
 - Retorna conjunto \mathbb{C} , $|\mathbb{C}| = k$ tal que

$$\forall x \in \mathbb{C}, y \in \mathbb{U} - \mathbb{C}, \delta(x, q) \leq \delta(y, q)$$

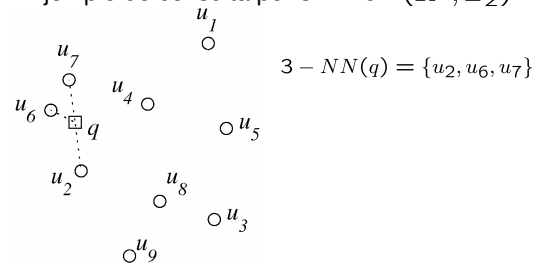
09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

20

Consultas por similitud

- Ejemplo de consulta por 3-NN en (\mathbb{R}^2, L_2)



09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

21

Efectividad y eficiencia

- Eficiencia
 - Se relaciona con el costo de búsqueda
 - Qué se mide: tiempo de CPU y tiempo de E/S
- Algoritmo ingenuo: búsqueda secuencial
- Estructuras de datos para agilizar búsquedas
 - Índices multidimensionales (*spatial access methods*)
 - Índices métricos (*metric access methods*)

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

22

Efectividad y eficiencia

- Eficacia
 - Calidad de la respuesta
 - En espacios vectoriales: objetos similares implica puntos cercanos
- Descriptores finos se obtienen con resoluciones altas
 - No necesariamente implican mejor eficacia

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

23

Evaluación de la efectividad

- Evaluación de la eficacia
 - Medir su habilidad de recuperar **objetos relevantes** de la BD y de evitar los objetos no relevantes
- Medidas de efectividad
 - "Ground truth": Colección de prueba
 - Medida de evaluación: cuantifica similitud entre objetos recuperados y objetos relevantes

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

24

Evaluación de la efectividad

- Objetos relevantes vs. no relevantes

	Deseado	No deseado
Encontrado	Positivo correcto (RP)	Falso positivo (FP)
No encontrado	Falso negativo (FN)	Negativo correcto (RN)

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

25

Evaluación de la efectividad

- Precisión y recall**
 - Precisión: ¿Cuántos de los objetos recuperados son relevantes?

$$Precision = \frac{RP}{RP+FP}$$

- Recall: ¿Cuántos de los objetos relevantes fueron encontrados?

$$Recall = \frac{RP}{RP+FN}$$

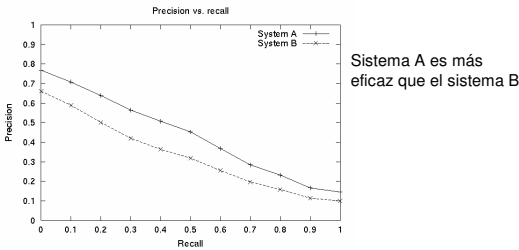
09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

26

Evaluación de la efectividad

- Gráfico precisión vs. recall



09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

27

Evaluación de la efectividad

- Medidas de un solo valor
 - R-precision** ("1st tier"): precisión cuando # objs recuperados = # objs relevantes
 - Bull-Eye Percentage** ("2nd tier"): recall cuando # objs recuperados = 2 × # objs relevantes

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

28

Colecciones de referencia

- Colección de referencia**
 - Colecciones usadas para probar modelos de RI y algoritmos
 - Incluye:
 - Conjunto de objetos
 - Conjunto de consultas
 - Conjunto de objetos relevantes a cada consulta

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

29

Colecciones de referencia

- Princeton Shape Benchmark**
 - Colección y herramientas para recuperación de modelos 3D
 - 1,814 modelos 3D:
 - Clasificación base para entrenamiento:
 - 90 clases, 907 modelos
 - Clasificación base para pruebas:
 - 92 clases, 907 modelos
 - URL: <http://shape.cs.princeton.edu/benchmark/index.cgi>

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

30

Agenda

- Tipos de datos multimedia
- Búsqueda por similitud
- Modelos de similitud para datos multimedia
- Temas de investigación

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

31

Búsqueda de imágenes

- Búsqueda por contenido de la imagen misma
 - Contenido derivado de la representación interna de la imagen (píxeles)
 - Evita los problemas de las anotaciones manuales
 - Características a considerar
 - Colores
 - Histogramas de colores
 - Texturas
 - Estructuras de segmentos de imágenes (madera, piedra, etc.)
 - Formas (contornos)
 - Modelo de similitud morfológico
 - Puntos de interés
 - SIFT

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

32

Búsqueda de imágenes

- De todas formas es un problema difícil...



09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

33

Histogramas de colores

- Histogramas de colores
 - Representación de la distribución de colores en una imagen
 - Definición del histograma de colores
 - Fijar el espacio de colores (e.g., RGB, HSV, CMY)
 - Elegir representantes de dicho espacio (*sample points*)
 - E.g., matriz en espacio de colores de $4 \times 4 \times 4 = 64$ colores o $8 \times 8 \times 8 = 512$ colores

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

34

Histogramas de colores

- Cálculo del histograma de colores
 - Por cada píxel, aumentar el contador de su representante más cercano en uno
 - Normalizar el histograma para hacerlo independiente del tamaño de la imagen
 - Ejemplos (64 representantes)



09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

35

Similitud basada en píxeles

- Concepto de la "imagen de diferencia" (tonalidades grises)

Consulta y resultados (64x64 píxeles)



Imagen de diferencia y distancia euclidiana



09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

36

SIFT

■ Puntos de interés



09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

37

SIFT

■ Match



09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

38

Ejemplo

■ Image Search:

- <http://prisma.dcc.uchile.cl/ImageSearch/>

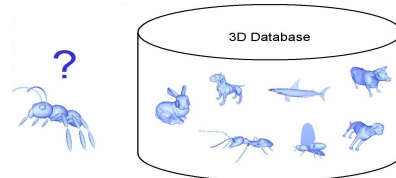
09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

39

Similitud de modelos 3D

■ Buscar modelos 3D similares (geometría)



09/11/2009

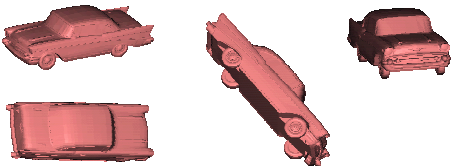
Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

40

Similitud de modelos 3D

■ Requerimientos para descriptores 3D

- Invariancia: rotaciones, traslaciones, escala



09/11/2009

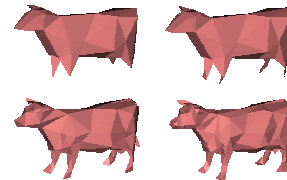
Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

41

Similitud de modelos 3D

■ Requerimientos para descriptores 3D

- Robustez con respecto al nivel de detalle



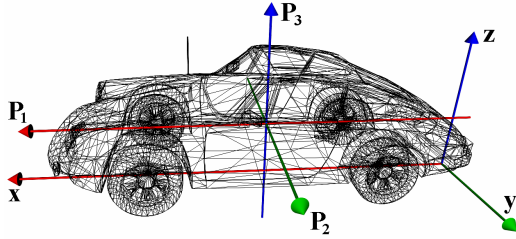
09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

42

Normalización

- Normalización con PCA



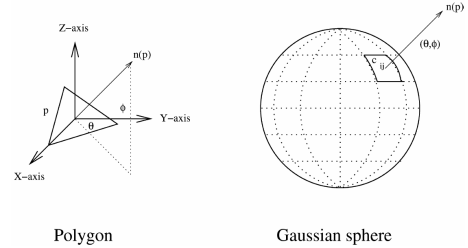
09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

43

Extended gaussian image

- Mapeo a esfera gaussiana



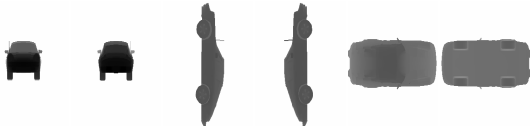
09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

44

Depth-buffer

- Basado en proyecciones 2D
- Usa información de profundidad



09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

45

Agenda

- Tipos de datos multimedia
- Búsqueda por similitud
- Modelos de similitud para datos multimedia
- **Temas de investigación**

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

46

Temas de investigación actual

- Análisis formal de algoritmos de búsqueda
- "Gap" semántico
- Búsqueda en espacios multi-métricos o no-métricos
- Aplicaciones para
 - Video (detección de copia)
 - Audio (identificar canciones)
 - Imágenes (búsqueda basada en *sketch*)

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

47

Fin

¡Gracias por su atención!



<http://prisma.dcc.uchile.cl>

09/11/2009

Benjamin Bustos - Búsqueda por Similitud en Bases de Datos Multimedia

48