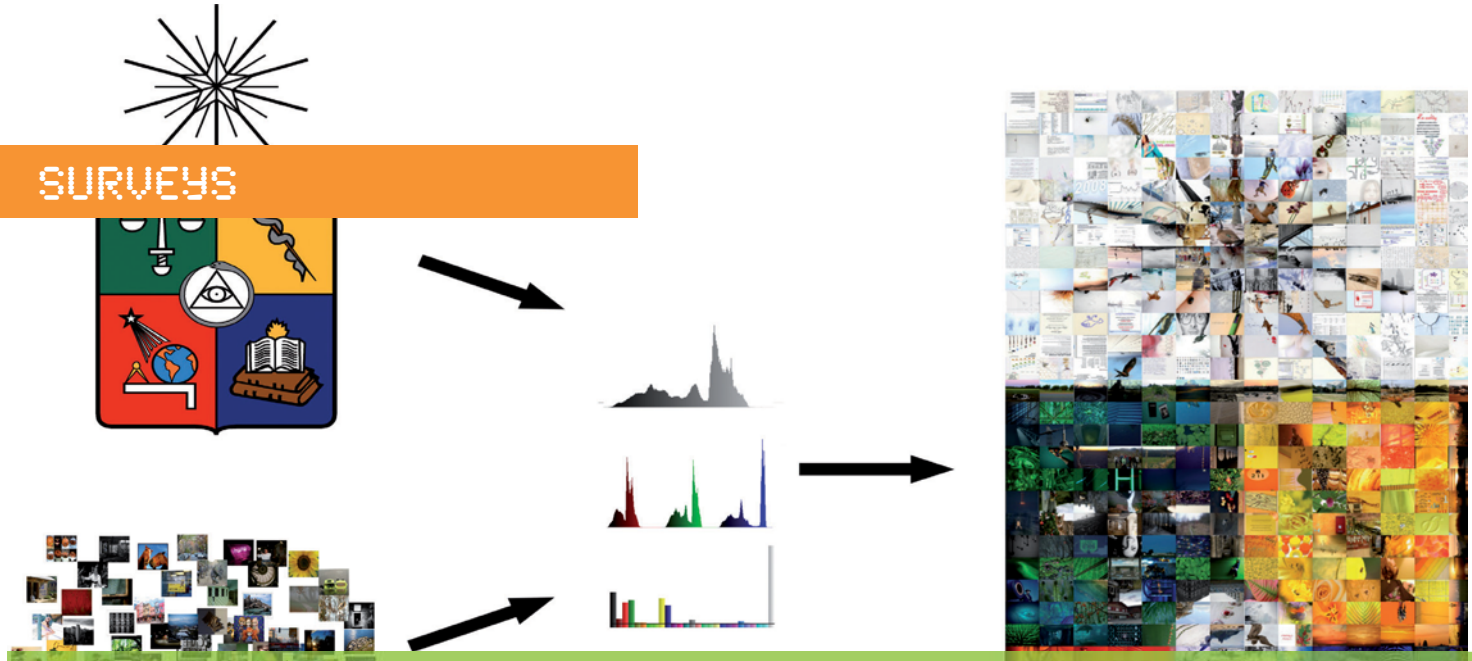
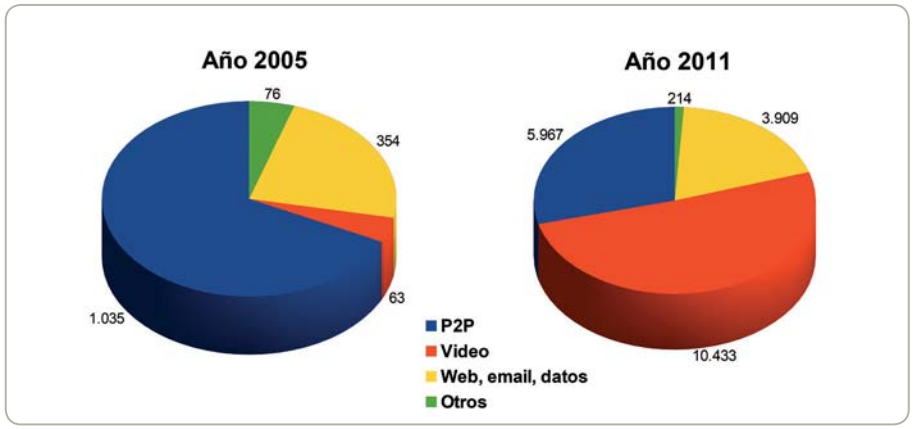


Búsquedas por contenido en imágenes y videos

El tráfico de datos en Internet ha aumentado exponencialmente en la última década. En 2005 las redes P2P dominaban las estadísticas del tráfico de datos de usuarios de Internet, mientras que la visualización de videos en línea recién comenzaba a masificarse con el lanzamiento de YouTube. Seis años después, el tráfico global de datos de usuarios ha aumentado más de doce veces, y específicamente la visualización de videos (esto es, reproducción de cortos de videos, visualización de películas, streaming de canales televisión, webcams públicas, etc.) corresponde a más de la mitad (ver Figura 1). Este crecimiento muestra un sorprendente aumento en el tráfico de datos, y en particular del impacto de los videos en el uso de ancho de banda.

Figura 1



Tráfico de datos de usuarios de Internet para 2005 y 2011 (fuente: [1] y [2]).

Juan Manuel Barrios
Doctor (c) en Ciencias mención Computación, Universidad de Chile. Ingeniero Civil en Computación y Magíster en Ciencias mención Computación Universidad de Chile. Director de Investigación Orand S.A.
jbarrios@dcc.uchile.cl.

Por otra parte, la investigación académica en tópicos de video se ha desarrollado desde hace varias décadas. Por ejemplo, la industria del software ha provisto de aplicaciones para edición de videos desde previo a este siglo. Así también, algunas de las técnicas que veremos más adelante provienen de las decadas de 1970 y 1980. Sin embargo, la actual masificación y universalidad en el uso de videos crea nuevas necesidades y presiona el desarrollo de nuevos y mejores algoritmos. Actualmente, muchos usuarios, investigadores y empresas estarán interesados en explorar posibilidades como administración, análisis, y gestión de grandes colecciones de videos, monitoreo y automatización de procesos en tiempo real, etc.

Multimedia Information Retrieval (MIR) es el área de investigación que tiene por objetivo la búsqueda y recuperación de información semántica desde documentos multimedia [3]. Un documento multimedia se puede entender como cualquier repositorio de información, ya sea estructurado o no. En particular, en este artículo nos enfocaremos en documentos audiovisuales, esto es imágenes, audio y videos, aunque un documento multimedia también comprende fuentes de información más genéricas como grafos, series de tiempo, páginas web, documentos XML, secuencias de ADN, etc. En general, se pueden destacar dos procesos fundamentales en cada sistema MIR: un proceso de *descripción de contenido* que calcula uno o más descriptores para cada documento, y un proceso de *búsqueda por similitud* que analiza la distribución de descriptores para encontrar descriptores parecidos de forma efectiva y eficiente.

Computer Vision (CV) es un área de investigación muy vinculada con MIR. Ambas áreas comprenden la adquisición y procesamiento de imágenes y videos, y la toma de decisiones según el análisis de descriptores. La principal diferencia proviene del contexto de uso: CV se enfoca mayormente en procesos en tiempo real, por ejemplo, detectar cuando un objeto específico aparece frente a la cámara de un robot; mientras que MIR se enfoca mayormente en técnicas de búsqueda y análisis de grandes colecciones, por ejemplo, buscar un objeto específico en imágenes de Internet.

PROCESAMIENTO DE IMÁGENES

Llamaremos procesamiento de imágenes a las técnicas que dada una imagen de entrada producen una imagen de mejor calidad de salida. El objetivo básico de estas técnicas es facilitar o mejorar la extracción de características.

Una imagen es una señal bidimensional discreta $I(x,y)=c$, donde cada celda (x,y) se denomina píxel ("picture element"). En imágenes grises c corresponde a un valor unidimensional con la intensidad del píxel, mientras que en imágenes en color c es un valor de (al menos) tres dimensiones. Cada dimensión de c se denomina canal, y la cantidad de valores posibles para representar cada canal se denomina profundidad. Comúnmente se usa profundidad 8 bits/canal, lo que permite representar hasta 2^8 tonos de gris o hasta $(2^8)^3 \approx 16$ millones de colores.

Operadores

Los operadores puntuales modifican el valor de cada píxel en forma independiente: dada una imagen de entrada I un operador puntual f produce una imagen de salida I′ definida como $I'(i,j)=f(I(i,j))$. Algunos ejemplos de operadores puntuales son el filtro de binarización ($f(x)=0$ si $x<t$, o 1 si $x\geq t$), corrección gamma ($f(x)=x^\gamma$) y ajuste de brillo y contraste ($f(x)=ax+b$).

La convolución es un operador lineal donde el nuevo valor de un píxel corresponde a una combinación lineal de los píxeles de su vecindad. A la ventana de ponderación se le conoce como filtro, máscara, o kernel. Existen diferentes filtros usados con diferentes objetivos. Por ejemplo, el efecto de desenfoque o *blur* se logra con la convolución de una imagen con un filtro promedio (ver Figura 2) o con un filtro gaussiano (en general, un filtro pasa bajo). Estos filtros descartan información de detalles y son usados para disminuir ruido o para reducir el tamaño de la imagen. Otros filtros comúnmente usados son Laplaciano (para resaltar detalles) y filtros de Roberts, Prewit y Sobel (para detección de bordes).

Algunos operadores no lineales comunes son el filtro de mediana, donde el nuevo valor es la mediana de los valores de la vecindad de cada píxel; filtro bilateral, el que es un filtro gaussiano con ponderaciones dinámicas por píxel; y filtros morfológicos, que marcan formas relevantes en la imagen según cómo se comportan al expandirse o contraerse al usar un elemento estructurante.

Detección de bordes

Los bordes corresponden a las zonas de la imagen donde existe un cambio abrupto en la tonalidad (ya sea de blanco a negro o de negro a blanco). Una técnica común para detectar bordes es el método del gradiente. En una función bidimensional el gradiente corresponde a un vector conteniendo las derivadas parciales según ambos ejes. La orientación del gradiente entrega la dirección de máximo cambio, mientras que su magnitud representa la intensidad de ese cambio. La convolución con filtros de Sobel permite aproximar las derivadas parciales, y estimar el valor del gradiente

Figura 2

a

$\frac{1}{9} \times$

1	1	1
1	1	1
1	1	1

b

c

a) Imagen original. b) Filtro promedio 3x3. c) Convolución con un filtro promedio.

Figura 3

-1	0	1	-1	-2	-1	0	1	0
-2	0	2	0	0	0	1	-4	1
-1	0	1	1	2	1	0	1	0

a b c

a) Filtros de Sobel ejes *x* e *y* . b) Filtro Laplaciano. c) Resultado de detección de Canny.

para cada píxel. Los píxeles de borde se definen como los puntos donde la magnitud del gradiente supera cierto valor umbral. Esta técnica tiende a dar bordes gruesos, los que pueden ser adelgazados usando un criterio basado en la segunda derivada. El Laplaciano se define como la suma de las segundas derivadas parciales de una función, y puede ser estimado por medio de una convolución con el filtro del mismo nombre. Luego, se pueden obtener bordes delgados si seleccionan los píxeles donde la magnitud del gradiente supera cierto umbral *t* y además el valor del Laplaciano es cero. Esta técnica fue formalizada por Canny [4], quien además incluye dos mejoras: realiza una selección incremental de puntos recorriendo en dirección perpendicular al gradiente (es decir, "caminando" por sobre el borde), e incluye un umbral incerteza t', donde un píxel cuya magnitud de gradiente está entre *t'* y *t* puede ser seleccionado como borde si es contiguo a un píxel que ya fue seleccionado como borde (ver Figura 3).

Una segunda técnica para detectar bordes es la Diferencia de Gaussianas (DoG). Ésta consiste en aplicar un filtro gaussiano y restar la imagen desenfocada con la original. Como el desenfoque afecta mayormente las zonas donde hay gran variación, al comparar el desenfoque con la imagen original las mayores diferencias se producen en los píxeles sobre el borde. Para reducir el nivel de ruido, se usa una imagen base I_1 con filtro gaussiano de desviaciones estándar σ, luego se calcula una imagen I_2 con un filtro gaussiano $k\sigma$ (para cierto paso k), y finalmente se aplica una binarización sobre la imagen de diferencia $I'(x,y) = |I_2(x,y)-I_1(x,y)|$ (ver Figura 4).

Figura 4

Resultado de bordes usando DoG.

En el caso de imágenes de objetos, el Análisis de Componentes Principales (PCA) permite obtener una imagen invariante a rotaciones. PCA busca un sistema de coordenadas donde los datos no tengan correlación lineal por medio de calcular valores y vectores propios de la matriz de covarianzas. En este caso, aplicando PCA sobre las coordenadas de los puntos de borde, con una rotación inversa de acuerdo al eje principal encontrado (esto es, el vector propio asociado al mayor valor propio) se puede normalizar la orientación de la imagen en función de su contenido (ver Figura 5). Cabe señalar que esta técnica no funciona correctamente con formas mayormente simétricas.

Figura 5

a) Imagen original. b) Uso de PCA para determinar los ejes principales.

Una vez seleccionados los píxeles de bordes, una tarea común es la de buscar diferentes estructuras que ellos forman, como rectas, circunferencias u otras estructuras paramétricas*. Dado un tipo de estructura a buscar (por ejemplo, rectas), existen dos técnicas comúnmente usadas para determinar la existencia de esa estructura y sus parámetros. Random Sample Consensus (RANSAC) es un algoritmo aleatorio que iterativamente crea y evalúa distintos candidatos para seleccionar el que obtuvo un mayor apoyo. Por ejemplo, en el caso de buscar rectas, se elige un par de puntos al azar, se calcula la ecuación de recta y se cuentan los puntos de borde que están sobre esa recta. Con suficientes intentos, se encontrará la recta que pasa sobre más puntos de borde. Sin embargo, cuando la probabilidad de encontrar un modelo correcto por selección aleatoria es muy baja (es decir cuando el modelo buscado es cumplido por muy pocos puntos) es recomendable preferir una técnica alternativa. La Transformada de Hough es un algoritmo exhaustivo y determinístico que reduce los parámetros posibles de la estructura a un conjunto finito. La ecuación de la recta se determina por dos parámetros, por tanto el espacio bidimensional de los parámetros se discretiza a una matriz de cierta dimensión fija. Para evitar problemas con los rangos de parámetros se debe usar un sistema de coordenadas polares. Cada punto de borde agrega un voto a los parámetros de todas las rectas que pasan por ese punto. La estructura buscada está definida por los parámetros que obtienen mayor votación.

DESCRIPCIÓN DEL CONTENIDO

La descripción del contenido consiste en analizar y resumir el contenido de cada documento creando uno o más descriptores o vectores característicos. Los descriptores de alto nivel representan características semánticas, como metadatos, tags o anotaciones, mientras que los de bajo nivel corresponden a estadísticas o patrones del contenido que pueden ser creados automáticamente, como promedios,

* Un caso particular es el de segmentación por contornos activos, tema que fue revisado en el número 5 de la Revista Bits de Ciencia.

varianzas e histogramas. Un tópico amplio de investigación es la generación automática de descriptores de alto nivel a partir de información de bajo nivel, es decir, predecir los conceptos que una persona asignaría a un documento analizando su contenido. Este problema está ligado a las áreas de machine learning, data mining, e inteligencia artificial. En el contexto de búsquedas por contenido, la diferencia que se produce entre los conceptos que un usuario asigna a un documento y los que un computador puede asignar automáticamente es conocida como "brecha semántica" (semantic gap) [31].

Descriptores globales de imágenes

Un descriptor global es un valor o vector que representa el contenido de toda la imagen. El nivel de parecido entre dos descriptores se mide por medio de una función de distancia o de disimilitud. Cuando la distancia entre dos descriptores es cercana a cero, es decir, cuando son espacialmente cercanos, se espera que las dos imágenes que produjeron esos descriptores sean parecidas entre sí. Una familia de distancias comúnmente usadas para comparar vectores son las distancias de Minkowski:

$$L_p(\vec{x},\vec{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}} \text{ para } \quad p \geq 1$$

En particular, p=2 corresponde a distancia euclidiana.

Un descriptor global muy simple consiste convertir la imagen a grises, escalarla a un tamaño fijo de *WxH* píxeles, y crear un vector de *W·H* dimensiones, donde cada valor corresponde a la intensidad gris de cada píxel escalado. Los descriptores pueden ser comparados con distancia euclidiana. Este descriptor permite comparar directamente pares de imágenes, sin embargo, el valor de la distancia es altamente afectado por ajustes simples de brillo o contraste.

Una variante común es usar la medición ordinal (ordinal measurement) [6], que consiste en reemplazar los valores gris de cada píxel por su posición relativa con respecto a los demás píxeles al ordenarlos de menor a mayor. Por ejemplo, si la imagen se reduce a 5x5 la medida ordinal es una permutación de los números 1 a 25 donde 1 corresponde al píxel más oscuro y 25 al más claro. Este descriptor es invariante a cambios en brillo y contraste, por lo que es usado para buscar imágenes duplicadas, sin embargo es altamente afectado por modificaciones parciales como inserción de subtítulos o logos.

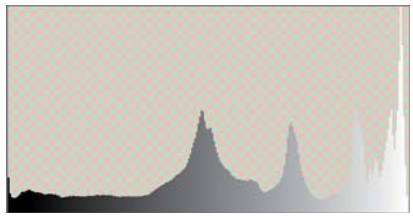
Otra variante es reducir la dimensionalidad de los descriptores usando PCA. Esto consiste en determinar los ejes principales, rotar los descriptores según estos ejes y luego descartar los ejes de menor varianza. Esta alternativa es especialmente útil cuando las imágenes de la colección tienen características comunes, por ejemplo una colección de rostros de personas. En este caso usando PCA se obtienen vectores propios que al ser combinados linealmente permiten obtener las imágenes originales de la colección. Esta es la idea base de la técnica conocida como *eigenobjects* para reconocer objetos, y que es aplicada en la autenticación de personas bajo el nombre de *eigenfaces* (Ver Figura 6).

Figura 6

Ejemplos de vectores propios (*eigenfaces*) de un conjunto de imágenes de rostros (fuente: prtools.org).

En una imagen en escala de grises, el histograma normalizado contiene la fracción de píxeles que tiene cada nivel de gris descartando su ubicación espacial. Dadas n observaciones (pixeles) y un conjunto de *k* categorías (rangos de intensidades) un histograma cuenta el número de observaciones que caen dentro de cada categoría o bin (ver Figura 7). El histograma puede ser visto como una distribución de probabilidad de que un píxel al azar tenga cierto valor de gris. Esta información es útil tanto para el procesamiento de imágenes como para representar el contenido. Un histograma es un vector de *k* dimensiones que puede ser comparado por distancias vectoriales o por test estadísticos como χ^2 o divergencia Kullback-Leibler.

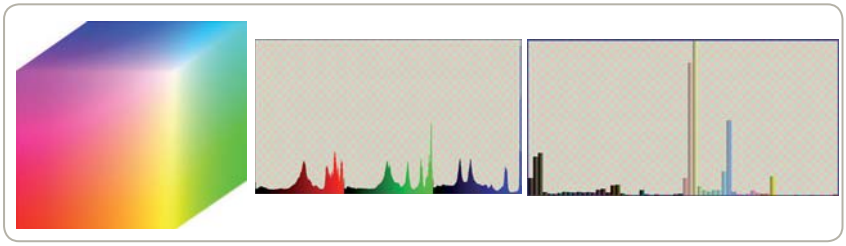
Figura 7



Histograma de grises para la imagen 2-a.

En el caso de histogramas de color existen diferentes enfoques. Un primer enfoque es calcular un histograma para cada canal en forma independiente, de esta forma el histograma de color es la concatenación de tres histograma de grises. Un segundo enfoque (y el más utilizado) es crear una partición regular fija del espacio de colores y calcular un histograma tridimensional. Por ejemplo, el espacio de colores RGB forma un cubo regular, dividiendo cada canal en cuatro rangos se obtendrán histogramas de $4^3=64$ bins (ver Figura 8). En este caso, la comparación de dos histogramas se puede realizar con distancia euclidiana, sin embargo, esta distancia no considera similitudes entre bins. La distancia Forma Cuadrática permite calcular distancias entre vectores usando una matriz de similitud entre dimensiones. Esta función permite ser más precisos en el cálculo de similitud entre histogramas, a un costo de más operaciones. La matriz de similitud se puede definir por medio de distancias de colores en el cubo RGB, sin embargo

Figura 8



Cubo RGB, histograma de color por canal, histograma de 27 colores de la imagen 2-A.

esto implica que la similitud entre blanco y verde es igual que entre blanco y azul, aún cuando perceptualmente la primera es más parecida que la segunda. Los espacios de color HSV, HLS y variantes corresponden a transformaciones geométricas de RGB. Se puede calcular histogramas directamente sobre estos espacios de colores, sin embargo en la práctica siguen estando fuertemente ligados a RGB. Una alternativa diferente es basarse en representaciones perceptuales del color, como las definidas por la Commission Internationale de L'Eclairage (CIE). Esta comisión desde los años veinte estudia la naturaleza física y perceptual del color, definiendo modelos basados en experimentos con observadores humanos. Durante los años setenta se definieron los espacios L*a*b* y L*u*v* que permiten calcular distancia entre colores asemejando la similitud percibida por un observador. Estos espacios de colores pueden ser usados para el cálculo de matrices de similitud entre colores.

Un tercer enfoque para crear histogramas de color consiste en utilizar una división dinámica del espacio de colores dependiendo de los colores en la imagen a representar. Esto es, determinar la mejor división y número de bins para representar fielmente el contenido de cada imagen. Esta división se logra por medio del análisis de la distribución de colores y/o por medio de clustering en el espacio de colores. Una vez creados los histogramas, su comparación se puede realizar con las funciones Earth Mover's Distance [7], que resuelve un problema de transporte entre bins según una función de costo, y Signature Quadratic Form [8], que extiende la forma cuadrática para incluir matrices de similitud intrahistogramas e interhistogramas.

Los bordes de la imagen entregan información valiosa sobre su contenido. Una alternativa consiste en representar los gradientes en la imagen por medio de un histograma de orientaciones [9]. Otra alternativa consiste en dividir la imagen en grupos de 2x2 píxeles y usar filtros de orientación para testear el tipo de borde [10] [11] (ver Figura 9). Una tercera alternativa es determinar bordes según el método de Canny y luego representar las ubicaciones dominantes de los bordes [9].

La información de texturas también puede ser usada para representar el contenido. Para esto, un análisis de frecuencias de la imagen revela patrones de textura en la imagen. Un descriptor consiste en obtener la transformada de Fourier de la energía y representar la energía de diferentes zonas del espacio de frecuencias de la imagen por medio de la energía obtenida por filtros Gabor [10]. Otra alternativa consiste en describir la imagen según el valor de los primeros coeficientes de la transformada discreta coseno [12]. Estos coeficientes tienen la propiedad que permiten reconstruir la imagen original a una aproximación ajustable según el número de coeficientes guardados.

Figura 9

1	-1	1	1	√2	0	0	√2	2	-2
1	-1	-1	-1	0	-√2	-√2	0	-2	2

Filtros de orientaciones de bordes.

Descriptores locales de imágenes

Un descriptor local representa sólo una pequeña zona de la imagen, y por tanto la imagen se representa por una cantidad variable de descriptores (del orden de cientos o miles para una sola imagen). Existen diferentes enfoques para decidir la ubicación y el tamaño de las zonas de interés en una imagen. Un enfoque es la detección de esquinas, que se basa en seleccionar zonas que mantienen una alta diferencia consigo misma bajo cualquier desplazamiento pequeño. Algunos algoritmos que usan esta idea son los detectores de Harris-Stephens [5], Shi-Tomasi y Harris-Laplace. Un segundo enfoque es la detección de manchas, que se basa en localizar zonas oscuras dentro de zonas claras (o zonas claras rodeadas de zonas oscuras). Estas zonas se pueden localizar por medio de detectores puntuales basados en DoG, o de formas arbitrarias llamadas Maximally Stable Extremal Regions (MSER). Otro enfoque es el llamado *dense sampling*, que consiste en utilizar una grilla densa de zonas de interés. Esta grilla permite obtener una gran cantidad de regiones, incluso en zonas donde los detectores anteriores detectan muy pocas.

Para cada zona de interés se calcula un descriptor del contenido en la zona. Los descriptores locales más usados son SIFT [13], SURF [14], y alguna de sus variaciones y extensiones. El descriptor SIFT divide la zona de interés en 4x4 regiones y calcula histogramas de orientaciones del gradiente, produciendo un vector de 128 dimensiones (ver Figura 10). Una variación común es PCA-SIFT [15] que reduce el largo de los vectores. SURF es un vector de 64 dimensiones que también representa las orientaciones del gradiente, pero usando la imagen integral para un procesamiento más rápido. SURF muestra casi los mismos resultados de efectividad que SIFT a un menor costo de procesamiento.

Figura 10

Descriptores SIFT sobre dos imágenes.

Para comparar dos imágenes I_1 e I_2, cada descriptor local de I_1 es comparado con cada descriptor local de I_2. En el caso de SIFT esta comparación se realiza con distancia euclidiana. Un descriptor *p* de I_1 se asocia con un descriptor *q* de I_2 cuando se cumplen dos condiciones: 1) *q* es el más cercano a *p* entre todos los descriptores de I_2 de acuerdo a la distancia entre descriptores; 2) la razón entre la distancia de *p* a *q* con la distancia de *p* al segundo más parecido en I_2 es menor a un parámetro *s* (usualmente 0.8). El primer criterio exige que se asocien zonas parecidas entre sí, mientras que el segundo criterio descarta calces que no sean suficientemente seguros o discriminativos (ver Figura 11).

Figura 11

Calces entre imágenes representadas por descriptores locales.

Una vez asociados descriptores locales similares, un proceso de correspondencia geométrica determina el mayor subconjunto de asociaciones que cumplen con una misma función de transformación espacial. Para esto, primero se debe definir un modelo de transformación geométrica a buscar. Los más comunes son escala+traslación, rotación+escala+traslación, transformación afín, y transformación de perspectiva u homográfica. Cada modelo requiere de una cantidad mínima de asociaciones para estar definido (por ejemplo, escala+traslación se define por dos asociaciones, mientras que una de perspectiva por cuatro). Luego se debe determinar el modelo de transformación geométrica que es cumplido por la mayor cantidad de asociaciones. En este caso los algoritmos RANSAC y Transformada de Hough ya descritos son aplicables (ver Figura 12).

Figura 12

Subconjunto de calces con coincidencia espacial usando RANSAC.

La correspondencia geométrica entrega la transformación que se debe llevar a cabo en la I_1 para que la mayor cantidad de zonas coincida con I_2. Esta información de postura puede ser aplicada en la composición de imágenes (llamado *image stitching*) para producir vistas panorámicas (ver Figura 13). También puede ser aplicada en realidad aumentada, donde objetos virtuales son transformados para ser incluidos en una escena real manteniendo coherencia con su entorno.

Descriptores de videos

Un video es la composición de una secuencia de imágenes del mismo tamaño más una pista de audio. La imágenes, que en este contexto se les llama cuadros o frames, se deben mostrar a cierta velocidad mínima (del orden de 24 frames por segundo) para percibir un movimiento fluido en vez de imágenes aisladas. Un shot se define como una secuencia de frames consecutivos de una misma cámara representando una acción continua en tiempo y espacio [16]. Los detectores de límites de shots más comunes comparan frames consecutivos, y reportan un cambio de shot cuando la diferencia entre dos frames supera un umbral. Algunas variaciones de esta técnica utiliza diferencias por zonas, diferencias de histogramas, ventanas temporales de comparación y uso de umbrales adaptativos [17].

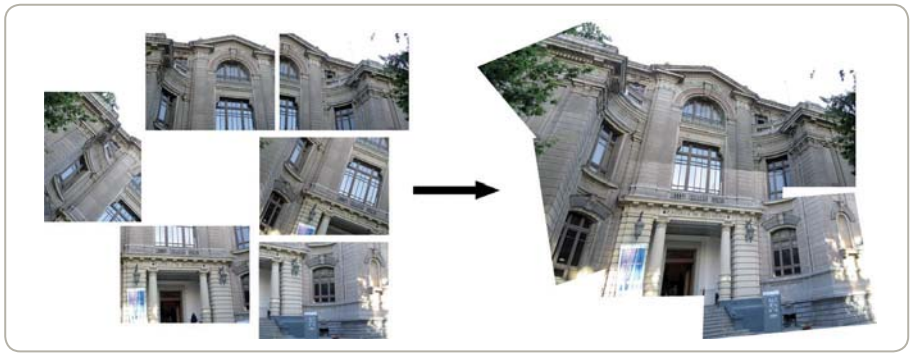
El flujo óptico consiste en calcular un vector de movimiento aparente entre frames consecutivos para cada píxel y/o grupos de píxeles. Es relativamente costoso de calcular, ya que requiere medir diferencias de frames consecutivos a distintas escalas y desplazamientos. El flujo óptico es usado para el tracking de objetos, codificación de videos y selección de keyframes.

Un keyframe es un frame que representa el contenido en una secuencia de video. En búsquedas por similitud, los keyframes permiten descartar frames redundantes y por tanto reducir la cantidad de información a describir. Los frames con mucho movimiento tienden a ser poco definidos, por tanto un criterio de selección es elegir frames mayormente estáticos. Para esto, una alternativa consiste en seleccionar los frames con menor diferencia con sus frames previos. Otra alternativa es usar el flujo óptico para seleccionar los frames cuyo largo total de vectores de movimiento sea mínimo. En el caso de querer seleccionar keyframes representativos para secuencias largas (por ejemplo, películas), se puede utilizar un algoritmo de clustering sobre frames para seleccionar los frames más cercanos a los centroides.

Para el cálculo de descriptores de videos, un primer enfoque conocido como hashing visual o video signature consiste en calcular un único descriptor para la secuencia completa. Este enfoque es útil para detectar videos duplicados con invariancia a pequeñas modificaciones visuales debidas a re-encoding o reducción de calidad [18]. Sin embargo, para poder detectar videos con segmentos comunes, se requiere calcular descriptores a un nivel más fino como keyframes o segmentos cortos.

La mayoría de los descriptores globales de imágenes pueden ser usados directamente sobre keyframes de videos. Una técnica común es calcular descriptores espacio-temporales para segmentos de videos por medio de agregación de descriptores de sus frames [19]. Otra técnica es calcular distancias temporales entre conjuntos de descriptores [20] [21].

Figura 13



Composición de imágenes al combinarlas según el modelo de transformación de perspectiva.

En el caso de descriptores locales para videos, una técnica consiste en calcular descriptores sobre keyframes y mantener los descriptores de las zonas de interés permanentes entre frames consecutivos [22]. Además, se pueden clasificar los tipos de zonas en estáticas o en movimiento [23]. También se ha realizado investigación para extender las técnicas de detección [24] y descripción [25] de zonas de interés para zonas espacio-temporales.

Con el objetivo de reducir la cantidad de descriptores locales producidos por un video y poder realizar búsquedas eficientes se desarrolló la técnica conocida como Bag-of-Visual-Words o Bag-of-Features [26]. Esta técnica se basa en usar un algoritmo de clustering sobre una gran cantidad de descriptores locales para determinar un conjunto discreto de N símbolos representativos, denominados vocabulario visual o codebook. Luego, los descriptores locales en una misma imagen o frame se resumen en un vector de dimensión N con las apariciones de cada símbolo. Las imágenes conteniendo cada símbolo se organizan en un índice invertido de N entradas. Luego, dada una imagen de consulta, el índice invertido permite determinar eficientemente todas las imágenes o frames que comparten un mismo símbolo. El índice invertido puede también contener información espacial de la aparición de cada símbolo, o en caso contrario, es necesario realizar un proceso de consistencia espacial entre imágenes o frames candidatos. Esta técnica ha sido foco de investigación durante los últimos años, abordando diferentes problemas como detección de videos duplicados [27], búsqueda de objetos en videos [28], clasificación de imágenes [29] y búsqueda de imágenes similares en la Web [30].

Una pista de audio es una señal unidimensional discreta $f(t)=a$, donde cada unidad de tiempo t se conoce como sample y el valor a corresponde a la amplitud de la onda de sonido. Una resolución de al menos ocho mil samples por segundo permite escuchar un sonido de calidad media mientras que 44 mil samples por segundo o más se considera de alta calidad. El análisis de audio se realiza por medio de ventanas pequeñas a intervalos regulares en la señal acústica. En vez de describir la señal directamente, los descriptores se basan en llevar la señal al dominio de las frecuencias (por medio de la Transformada de Fourier) y describir la energía de las frecuencias audibles. El descriptor más usado para audio es el Mel-frequency cepstral coefficients (MFCC) [32]. Otros descriptores son el de Philips [33], chroma features [34] y un descriptor enfocado en la duplicidad de pistas de audio [35].

Una vez descrito cada video por medio de descriptores globales, locales y/o acústicos ya sea por keyframes o shots, la comparación entre videos debe considerar todas estas variables. Las técnicas conocidas como *late fusion* [36] dividen la comparación en subsistemas independientes para cada tipo de descriptor y el resultado de cada uno se une ya sea por unión, intersección o agregación de scores. Por otra parte, las técnicas conocidas como early fusion intentan realizar una comparación combinada entre frames usando todos los descriptores en una única función de distancia [35].

Finalmente, una comparación de videos basada en keyframes requiere de un proceso de correlación espacio-temporal entre videos candidatos. La comparación entre dos videos puede ser modelada por un grafo bipartito donde los keyframes son nodos y la similitud entre keyframes son pesos de aristas. La similitud global de dos videos se puede determinar mediante un algoritmo de flujo máximo [37]. Otra alternativa es extender los modelos de coherencia espacial a modelos de coherencia espacio-temporal [38].

APLICACIONES

Existen diferentes aplicaciones para las técnicas de análisis y búsqueda de imágenes y videos. Por ejemplo, mediante análisis en tiempo real se pueden lograr usos industriales como inspección automática de control de calidad, autenticación biométrica, monitoreo y detección de eventos en cámaras de seguridad, o interfaces humano-computador a través de cámaras, etc. Mediante análisis de grandes volúmenes de imágenes y videos se pueden resolver problemas como organización automática de colecciones multimedia, detección de imágenes y videos similares, descubrimiento de imágenes o videos relacionados, detección de falsificación de imágenes, asignación automática de etiquetas semánticas, reconocimiento de eventos multimedia, etc. Además, existen subáreas especializadas para distintos dominios como el análisis de imágenes médicas, satelitales y astronómicas.

En el caso específico de videos, TRECVID es una conferencia organizada anualmente por el National Institute of Standards and Technology (NIST) de EE.UU que fomenta la investigación en la recuperación de información en videos [39]. Esta conferencia presenta problemas o desafíos relacionados con el análisis de videos, publica colecciones de referencia, y evalúa la participación de equipos que se inscriben libremente. Se han realizado evaluaciones para la detección de límites de shots, detección de copias de videos, búsquedas de objetos conocidos, monitoreo de cámaras de seguridad, asignación de etiquetas y detección de eventos multimedia. En estas evaluaciones es común ver equipos de universidades y empresas de diferentes partes del mundo, lo que prueba que el análisis de documentos multimedia es un tema desafiante que requiere de la participación de equipos multidisciplinarios, y la unión de la industria con la academia prueba la relevancia académica del área así como su proyección económica.

Un equipo del DCC de la Universidad de Chile participó en TRECVID en la evaluación de detección de copias de videos durante 2010 y 2011. Los resultados fueron satisfactorios al ser comparados con otros equipos y sistemas del estado del arte [20] [35]. El software desarrollado durante esa participación, llamado P-VCD, ha sido liberado como Open Source con licencia GPL. Actualmente, la empresa Orand ha apoyado el desarrollo de este software, con el que se ha participado en la evaluación de búsqueda de objetos (Instance Search) en TRECVID 2012 [40].BITS

REFERENCIAS

[1] Cisco Systems Inc. Global IP Traffic Forecast and Methodology, 2006–2011. White paper, 2007.

[2] Cisco Systems Inc. Cisco Visual Networking Index: Forecast and Methodology, 2011-2016. White paper, 2012.

[3] M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications and Applications, 2(1):1-19, 2006.

[4] Canny, J., A Computational Approach To Edge Detection. IEEE Trans. Pattern Analysis and Machine Intelligence, 8(6):679–698, 1986.

[5] C. Harris and M. Stephens. A combined corner and edge detector. In Proc. of the Alvey Vision Conference, 147-151. The Plessey Company, 1988.

[6] C. Kim. Content-based image copy detection. Signal Processing: Image Communication, 18(3):169-184, 2003.

[7] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision, 40(2):99-121, 2000.

[8] C. Beecks, M. S. Uysal, and T. Seidl. Signature Quadratic Form Distance. In Proc. of ACM Int. Conf. on Image and Video Retrieval (CIVR), 438-445, 2010.

[9] A. Hampapur and R. Bolle. Comparison of distance measures for video copy detection. In Proc. of the IEEE int. conf. on Multimedia and Expo (ICME), 737-740. IEEE, 2001.

[10] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. IEEE Transactions on Circuits and Systems for Video Technology, 11(6):703-715, 2001.

[11] K. Iwamoto, E. Kasutani, and A. Yamada. Image signature robust to caption superimposition for video sequence identification. In Proc. of the int. conf. on Image Processing (ICIP), 3185-3188. IEEE, 2006.

[12] X. Naturel and P. Gros. A fast shot matching strategy for detecting duplicate sequences in a television stream. In Proc. of the int. workshop on Computer Vision meets Databases (CVDB), 21-27. ACM, 2005.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2):91-110, 2004.

[14] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Speeded-up robust features (SURF). Computer Vision and Image Understanding, 110(3):346-359, 2008.

[15] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In Proc. of the intl. conf. on Computer Vision and Pattern Recognition (CVPR), II-506-513. IEEE, 2004.

[16] A. Hanjalic. Shot-boundary detection: unraveled and resolved? IEEE Transactions on Circuits and Systems for Video Technology, 12(2):90-105, 2002.

[17] J. S. Boreczky and L. A. Rowe. Comparison of video shot boundary detection techniques. Journal of Electronic Imaging, 5(2):122-128, 1996.

[18] B. Coskun, B. Sankur, and N. Memon. Spatio-temporal transform based video hashing. IEEE Transactions on Multimedia, 8(6):1190-1208, 2006.

[19] X. Wu, A. G. Hauptmann, and C.-W. Ngo. Practical elimination of near-duplicates from web video search. In Proc. of the int. conf. on Multimedia (ACMMM), 218-227. ACM, 2007.

[20] J. M. Barrios and B. Bustos. Competitive content-based video copy detection using global descriptors. Multimedia Tools and Applications, Springer, 2012.

[21] C. Kim and B. Vasudev. Spatiotemporal sequence matching for efficient video copy detection. IEEE Transactions on Circuits and Systems for Video Technology, 15(1):127-132, 2005.

[22] X. Anguera, T. Adamek, D. Xu, and J. M. Barrios. Telefonica research at trecvid 2011 content-based copy detection. In Proc. of TRECVID. NIST, 2011.

[23] J. Law-To, O. Buisson, V. Gouet-Brunet, and N. Boujemaa. Robust voting algorithm based on labels of behavior for video copy detection. In Proc. of the int. conf. on Multimedia (ACMMM), 835-844. ACM, 2006.

[24] G. Willems, T. Tuytelaars, and L. V. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In Proc. of the european conf. on Computer Vision (ECCV), 650-663. Springer, 2008.

[25] G. Willems, T. Tuytelaars, and L. V. Gool. Spatio-temporal features for robust content-based video copy detection. In Proc. of the int. conf. on Multimedia Information Retrieval (MIR), 283-290. ACM, 2008.

[26] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In Proc. of the IEEE int. conf. on Computer Vision (ICCV), 1470-1477. IEEE, 2003.

[27] M. Douze, A. Gaidon, H. Jegou, M. Marsza lek, and C. Schmid. Inria lear's video copy detection system. In Proc. of TRECVID. NIST, 2008.

[28] D.-D. Le, C.-Z. Zhu, S. Poullot, V. Q. Lam, D. A. Duong, and S. Satoh. National institute of informatics, japan at trecvid 2011. In Proc. of TRECVID. NIST, 2011.

[29] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluation of color descriptors for object and scene recognition. In Proc. of the intl. conf. on Computer Vision and Pattern Recognition (CVPR), 1-8. IEEE, 2008.

[30] H. Jegou, M. Douze, and C. Schmid. Packing bag-of-features. In Proc. of the IEEE int. conf. on Computer Vision (ICCV), 2357-2364. IEEE, 2009.

[31] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12):1349-1380, 2000.

[32] F. Zheng, G. Zhang, and Z. Song. Comparison of Different Implementations of MFCC. Journal of Computer Science & Technology, 16(6): 582–589, 2001.

[33] J. Haitsma and T. Kalker. A highly robust audio fingerprinting system. In Proc. of the int. symp. on Music Information Retrieval (ISMIR), 2002.

[34] D. P. W. Ellis. Classifying music audio with timbral and chroma features. In Proc. of the int. symp. on Music Information Retrieval (ISMIR), 2007.

[35] J. M. Barrios, B. Bustos, and X. Anguera. Combining features at search time: Prisma at video copy detection task. In Proc. of TRECVID. NIST, 2011.

[36] C. G. Snoek, M. Worring, and A. W. Smeulders. Early versus late fusion in semantic video analysis. In Proc. of the int. conf. on Multimedia (ACMMM), 399-402. ACM, 2005.

[37] H.-K. Tan, C.-W. Ngo, R. Hong, and T.-S. Chua. Scalable detection of partial near-duplicate videos by visual-temporal consistency. In Proc. of the int. conf. on Multimedia (ACMMM), 145-154. ACM, 2009.

[38] A. Joly, O. Buisson, and C. Frelicot. Content-based copy retrieval using distortion-based probabilistic similarity search. IEEE Transactions on Multimedia, 9(2):293-306, 2007.

[39] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In Proc. of the int. workshop on Multimedia Information Retrieval (MIR), 321-330. ACM, 2006.

[40] J. M. Barrios and B. Bustos. PRISMA-ORAND: Instance Search Based on Parallel Approximate Searches. In Proc. of TRECVID. NIST, 2012