UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

PROYECTO FIN DE CARRERA

Estudio comparativo de descriptores visuales para la detección de escenas cuasi-duplicadas

Óscar Boullosa García
Febrero de 2011

Estudio comparativo de descriptores visuales para la detección de escenas cuasi-duplicadas

AUTOR: Óscar Boullosa García
TUTOR: Víctor Valdés López
PONENTE: José María Martínez Sánchez

Video Processing and Understanding Lab

Video Processing and Understanding Lab
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Febrero de 2011

Abstract

This M.Sc. Thesis presents a comparative study of visual descriptors with an exhaustive evaluation for near-duplicate[1] scene detection within video content. The concept of near-duplicate images refers to a different geometric and photometric transformations of the images that belong to the same scene among several types of video scene. As a content set for the study we have created a proprietary database of real images made up from different sets of videos such as the *BBC rushes videos* from *TRECVID Video Retrieval Evaluation*[2] or *Time-Slice® Film's videos from Vimeo*[3] among others.

We first start by describing certain features, techniques and methods of duplicate image detection based on different algorithms and image features proposed in the literature. After an introduction to near-duplicate scene detection and an exposition of different applications related to this issue, a review of the state of the art of the selected visual descriptors and content-based image retrieval techniques is presented. Below a comparative analysis of the selected methods is performed, including the proposal of improvements for the existing techniques by their combination.

In the experimental section, we have carried out different improvements of every basic algorithm which are included in a optimization stage showing the advantages and disadvantages of each visual descriptor by a comparative evaluation. Then individual and combined refinements have been performed in order to demonstrate the improvements sought respect to the original results. Besides, a computational cost evaluation of the improved techniques has been performed in order to obtain more rigorous and useful conclusions for the inclusion of the techniques to in real systems.

As a result we have obtained the combination of techniques that perform better in terms of precision and computational performance. Finally, the conclusions and feature work directions are presented.

Keywords

Visual descriptors, low-level features, histogram, color layout, correlogram, SIFT, SURF, near-duplicate detection, video summarization, video retrieval.

[1] From this point the near-duplicate image term will be referenced as duplicate image
[2] http://trecvid.nist.gov/
[3] http://vimeo.com/timeslice

Resumen

Este Proyecto presenta un estudio comparativo en profundidad de diversos descriptores de imagen con una exhaustiva evaluación de los mismos dentro del marco de la detección automática de imágenes afectadas por diferentes trasformaciones tanto geométricas como fotométricas y pertenecientes a distintos tipos de escenas. Como conjunto de datos para la implementación se ha creado una base de datos propia conformada a partir de diferentes colecciones de vídeos, entre las que cabe mencionar los *BBC rushes videos* pertenecientes al *TRECVID Video Retrieval Evaluation*[4] o los vídeos contenidos en *Time-Slice® Film's videos from Vimeo*[5] entre otros.

Como punto de inicio, se parte de una serie de características, técnicas y métodos de detección de imagenes cuasi-duplicadas[6] basados en diferentes algoritmos y peculiaridades de las imágenes propuestos en la literatura. Tras una introducción a la detección de imágenes cuasi-duplicadas pertenecientes a una misma escena y hacer referencia a algunas de las distintas aplicaciones que guardan relación con esta problemática, se presenta un estudio del estado del arte sobre las diferentes técnicas o descriptores de imagen elegidos para el desarrollo de este proyecto para a continuación realizar un análisis comparativo de los métodos elegidos, incluyendo las propuestas de mejoras en las técnicas actuales mediante mejoras parciales en las mismas así como combinación de algunas de ellas.

En la sección experimental se llevan a cabo diversas pruebas relacionadas con la optimización de los algoritmos base de cada uno de las descriptores seleccionados de modo que se puedan evidenciar de forma comparativa las ventajas e inconvenientes de cada uno de ellos. A continuación se elaboran diferentes refinamientos para que tanto de manera individual como combinada se puedan presentar mejoras respecto de los resultados iniciales. Además de la evaluación de las mejoras funcionales, también se realiza un análisis del coste computacional de las mismas elaborando de este modo unas conclusiones más rigurosas con el fin de poder ser utilizadas en una implantación dentro de un sistema real.

Como resultado global se han obtenido una serie de combinaciones de técnicas que aumentan el rendimiento del sistema. Finalmente, se presentan las conclusiones y se proponen líneas de trabajo futuro.

Palabras clave

Descriptores de imagen, características de bajo nivel, histograma, color layout, correlograma, SIFT, SURF, detección de imágenes cuasi-duplicadas, resúmenes de vídeo, recuperación de vídeos.

[4] http://trecvid.nist.gov/

[5] http://vimeo.com/timeslice

[6] En adelante el término imágenes cuasi-duplicadas será referenciado como imágenes duplicadas

Agradecimientos

Tras un largo camino recorrido es el momento de hacer balance y agradecer de la manera más sincera a todas aquellas personas que han contribuido de alguna manera en la realización de este Proyecto.

En primer lugar quiero agradecer a mi ponente José María Martínez por darme la oportunidad de realizar este proyecto y por haber formado parte de un grupo de profesores que han incentivado en mi el interés por aprender y las ganas de trabajar durante todos estos años.

Quiero agradecer muy especialmente a mi tutor Víctor Valdés por la dedicación y disponibilidad en todo momento que aderezadas con su experiencia han hecho posible la elaboración de este Proyecto. Gracias por tener siempre un hueco en el que poder discutir juntos las dudas surgidas y por los consejos aportados para llevar el trabajo aquí presentado un paso más allá y sugerir la creación de su publicación. En resumen, un gran profesor que disfrutarán los estudiantes de esta escuela.

Gracias a mis compañeros de universidad por haber formado parte de mi vida durante una larga etapa que siempre recordaré y en la que sin ellos no hubiera sido lo mismo. Gracias a Pablo, Fabio y Luis por todos los momentos de risas, prácticas y abrazos que al paso del tiempo se han transformado en un gran cariño. A Pablo, Javier, Rubén y compañía por las risas y los buenos momentos durante las clases y los descansos. También a José Rubén por acompañarme en el camino que juntos emprendimos dejando atrás a otros buenos amigos y aguantar mis diferencias durante todos estos años.

Gracias a mi familia por todo el apoyo y la ilusión compartida durante toda mi vida. Vuestro amor y cariño me han ayudado a superar los malos momentos y disfrutar de todos los buenos. Son los valores que me habéis transmitido los que me han enseñado lo que realmente es importante en la vida. Gracias de verdad por todo vuestro sacrificio para permitirme haber estudiado esta carrera.

Y en especial, gracias a Catalina, la persona más importante de mi vida por estar a mi lado y compartir todos los momentos. Tus ánimos y amor incondicional me han aportado la fuerza suficiente para superar cualquier adversidad. Nunca podre agradecer todos los esfuerzos que has hecho por mi para conseguir que sea, por encima de todo, una persona feliz y podamos seguir nuestro camino juntos, compartiendo toda la grandeza del mañana. Te quiero.

Óscar Boullosa García,
Febrero 2011.

Índice General

Índice de figuras

Índice de tablas

1. Introducción

1.1. Motivación del proyecto

En la última década se ha producido un aumento sin precedentes con respecto a la cantidad de contenidos audiovisuales disponibles debido principalmente al uso masivo de Internet y a la proliferación de dispositivos multimedia en el ámbito cotidiano tanto a nivel empresarial como personal. La necesidad de soluciones adecuadas es cada vez más demandada en distintas y tan variadas áreas como Internet, aplicaciones de usuario, TV, bibliotecas digitales, aplicaciones medicas, etc. y se requieren métodos de acceso y gestión de la información para hacerla disponible de una manera más eficiente. Se pueden mencionar como ejemplos grandes bases de datos de vídeo como Youtube, distribuidores de contenidos que desean crear de forma automática resúmenes de vídeos o agencias de noticias y compañías de radiodifusión que contienen grandes colecciones de vídeos y que podrían crear pequeños resúmenes para facilitar las búsquedas por contenido de una manera más eficiente.

Las técnicas mencionadas anteriormente tienen una característica en común: representan el contenido mediante valores numéricos, que componen las características o descriptores, y que extraen diferentes propiedades del contenido permitiendo así un tratamiento más objetivo y con independencia de la naturaleza del contenido. Con respecto al ámbito visual, dichas técnicas se basan, por lo general, en comparaciones visuales centradas en la eliminación de información redundante o la recuperación de segmentos de vídeo similares. Las comparaciones visuales se llevan a cabo haciendo uso de diferentes los descriptores visuales existentes. Sin embargo, tales descriptores han sido evaluados normalmente centrándose en su rendimiento respecto de la identificación o recuperación de contenidos similares desde un punto de vista semántico y no focalizando en la capacidad para la detección de imágenes pertenecientes a una misma escena o pequeños cambios en las mismas. Este proyecto se centra en la evaluación de los descriptores de imagen respecto de este tipo de situaciones: la detección de escenas similares afectadas por diferentes transformaciones tales como cambios de iluminación, variaciones de zoom, cambios del punto de vista del objetivo y movimiento de objetos en la escena.

Se han llevado a cabo diferentes trabajos e investigaciones respecto a la evaluación de descriptores en diversas áreas como el reconocimiento de objetos [1] o características particulares [2], la detección de copias en los contenidos audiovisuales [3] o la recuperación de imágenes basada en el contenido [4], más conocido por el término anglosajón, Content-Based Image Retrieval (CBIR). Las evaluaciones mencionadas y otras relacionadas han sido desarrolladas en un contexto de comparación individual, sin embargo no se ha llevado a cabo una exhaustiva evaluación comparativa de descriptores visuales respecto de la transformaciones anteriormente comentadas en el seno de escenas duplicadas.

La cuestión principal en la que se ahonda en este proyecto es: ¿Que características visuales resultan más precisas en cuanto a la representación del contenido

y alcanzan un mayor rendimiento con respecto a la detección de imágenes cuasi-duplicadas y pertenecientes a una misma escena? Esta cuestión es ampliamente investigada examinando el comportamiento de un conjunto representativo de los diferentes descriptores de imagen. Esta tarea sobre cómo de bien se comportan las diferentes características de las imagenes está íntimamente relacionada con la cuestión sobre qué características pueden ser combinadas para obtener mejores resultados en tareas concretas. Tomando parte en esta última cuestión se elaboran diferentes métodos de combinación de características basados en la correlación de las características individuales.

Para la evaluación de las diferentes características o descriptores se ha utilizado una base de datos propia de imágenes reales que ha sido creada a partir de un muestreo manual de diferentes colecciones de vídeos y que conforma un buen punto de partida para evaluar el rendimiento de los descriptores seleccionados así como de las nuevas combinaciones desarrolladas.

1.2. Objetivos y enfoque

Este proyecto tiene como principal objetivo presentar un estudio comparativo sobre distintos descriptores de imagen así como sobre las diferentes distancias empleadas con la finalidad de analizar la robustez y fragilidades de cada uno de ellos en diferentes situaciones. Tras el análisis, se lleva a cabo un proceso de refinamiento y combinación de los descriptores con el propósito de mejorar de los resultados iniciales con respecto a la tarea de la detección de imágenes relativas a escenas cuasi-duplicadas.

Con este objetivo nace la necesidad de establecer un marco comparativo de referencia para evaluar los diferentes resultados obtenidos a lo largo de todo el proyecto. Como primer paso se lleva a cabo la tarea de confeccionar una base de datos de imágenes y que será utilizada durante las diferentes etapas de optimización, comparación y combinación de descriptores de las que se compone este estudio comparativo. Al mismo tiempo se realiza un proceso de selección de los algoritmos que representarán las descripciones matemáticas que se encuentran detrás de las características de los descriptores de imagen elegidos.

La recopilación de las imágenes se ha llevado a cabo teniendo en cuenta cuatro transformaciones básicas en los que se centrará nuestro trabajo sobre la detección de duplicados como son: cambios de iluminación, posición angular, movimiento dentro de la imagen y zoom. Por otro lado la elección de los descriptores de nuestro estudio esta basada en una gran tarea de estudio del arte sobre los descriptores de imagen con la consiguiente elección de un grupo de cinco de ellos con los que se intenta abarcar distintas modalidades referentes a la complejidad, popularidad, usabilidad en aplicaciones reales y posibilidades de combinación entre ellos. En base a estos criterios se han elegido los siguientes descriptores de imagen: *histograma de color, correlograma, color layout (perteneciente al estándar MPEG-7), Scale Invariant Feature Transform (SIFT) y Speed Up Robust Features (SURF).*

Una vez seleccionados los descriptores de imagen, se implementarán los algoritmos correspondientes a cada uno de ellos sobre Matlab® que permitan caracterizar el contenido de las imágenes prestando importancia a la parametrización de algunos de ellos para la consecución de nuestro objetivo. Realizaremos refinamientos y cambios progresivos sobre los descriptores con el animo de obtener mejores resultados para finalmente y mediante combinaciones de los descriptores elevar las prestaciones de los mismos cubriendo así ciertas debilidades individuales.

Para ello se ha dividido la metodología de trabajo en 4 bloques que se detallan a continuación:

Familiarización: En la primera etapa del presente proyecto se ha llevado a cabo una primera toma de contacto con los diferentes aspectos relacionados con la detección de duplicados de imagen y las particularidades que relacionan este tipo de imágenes dentro de una misma escena. La búsqueda y lectura de diferentes documentos científicos han permitido la elaboración y presentación de una síntesis sobre la problemática mencionada.

Investigación: Una vez han sido identificados las particularidades que afectan a la problemática mencionada se ha realizado un estudio sobre el estado del arte en profundidad referente a los diferentes descriptores de imagen y distintas distancias para las comparaciones elegidas para la realización de este proyecto. Esta etapa aporta el conocimiento necesario para la identificación y descubrimiento de posibles nuevas aportaciones sobre los trabajos iniciales de cara a la implementación final.

Implementación y desarrollo: La tarea principal es la de crear un marco comparativo para el estudio de diferentes descriptores de imagen. Este marco esta compuesto tanto de una base de datos de imágenes común durante todo el proyecto como de y de esta manera elaborar diferentes combinaciones de los mismos con los consecuentes resultados y que éstos puedan ser comparados de manera objetiva con los resultados previos respecto de la cuestión central de este proyecto.

Estudio de resultados y formalización de las conclusiones: Finalmente se realiza un estudio exhaustivo de los resultados obtenidos así como una elaboración de las conclusiones al mismo tiempo que se proponen nuevas mejoras del sistema y lineas de trabajo futuro.

Escritura del proyecto: Si bien esta tarea se ha desarrollado de forma continua a lo largo de todo el periodo del proyecto, gran parte de su elaboración se ha realizado tras la conclusión le las etapas de desarrollo y estudio de resultados.

1.3. Organización de la memoria

La memoria del presente proyecto se estructura en una serie de capítulos cuyo idea principal y contenidos se exponen a continuación:

El Capítulo 1 contiene la introducción, la motivación, los objetivos y enfoque así como la organización del Proyecto Fin de Carrera.

En el Capítulo 2 se expone la problemática relacionada con la detección de imágenes duplicadas representantes de una misma escena. Es aquí donde se define ampliamente el concepto de descriptores de imagen, se presentan distintas clasificaciones de los descriptores y se hace referencia a distintas aplicaciones que comparten la misma problemática.

Tras la exposición de la problemática mencionada se exponen los diferentes criterios de selección de los descriptores utilizados en este proyecto en el Capítulo 3. Así mismo se realiza un estudio más en profundidad sobre el estado del arte de los descriptores visuales seleccionados destacando las ventajas e inconvenientes de cada uno de ellos.

En el Capítulo 4 se detalla la creación de una base de datos de imágenes propia así como los diferentes sistemas de evaluación que se utilizarán para obtener los resultados. Serán las diferentes etapas de optimización, comparación y combinación de descriptores la parte central de este capítulo y en las que se detallarán por un lado los distintos parámetros utilizados tanto en la implementación de los algoritmos base como en las mejoras realizadas sobre el estado inicial y por otro la presentación de los resultados obtenidos en cada una de ellas. Finalmente se llevará a cabo un estudio computacional de los descriptores con el objetivo de elaborar conclusiones más objetivas para la utilización de estos descriptores en una aplicación real.

El Capítulo 5 contiene las conclusiones extraídas del trabajo y resultados obtenidos así como posibles lineas de trabajo futuro.

Las referencias consultadas para la elaboración de este Proyecto pueden encontrarse al final de esta memoria en la sección de Bibliografía seguida por una serie de Anexos en los que figuran diferentes ejemplos de las imágenes devueltas por cada descriptor así como el presupuesto y el pliego de condiciones.

2. Descriptores de Imagen

2.1. Introducción a los Descriptores de Imagen

La evolución tecnológica de los sistemas de comunicación y el grado de madurez que han alcanzado áreas tan diferentes como el procesamiento de señal, las bases de datos, el tratamiento multimedia y en gran medida el desarrollo y uso masivo de Internet, han contribuido a la inundación de información audiovisual en formato digital en cantidades desproporcionadas. Unido a esto podemos mencionar al caso, las colecciones de imágenes de ámbito privado, ya sea a nivel de usuario o como parte de una organización empresarial, que han originado la demanda de sistemas capaces de realizar una gestión y almacenamiento de la información de forma eficiente.

Esta gestión puede verse como el resultado final del proceso, si bien en el mismo preceden conceptos y etapas que permiten abordar la tarea, entre otras, de extraer información relevante de las imagenes para que éstas puedan ser procesadas de manera eficiente; esta tarea viene a consistir en intentar describir el contenido de los distintos tipos de información multimedia para efectuar la gestión posterior de acuerdo a la finalidad del sistema.

El contenido de una imagen esta codificado digitalmente en el valor de cada una de las unidades mínimas de información que la componen llamadas pixels. De esta manera los pixels representan el nexo de unión entre el contenido abstracto de sus valores y las características propias de una imagen que entendemos como relevantes para el humano. Es por ello que cualquier método de gestión de imagenes basado en su contenido deberá guardar algún tipo de relación o actuar sobre el valor de los mismos.

Respecto de la necesidad de describir el contenido de la información multimedia de forma objetiva y automatizada, surgen como respuesta los descriptores audio-visuales. Más concretamente en el caso de las imágenes podemos referirnos a los descriptores de imagen.

Idealmente, un descriptor visual debería poseer las siguientes propiedades:

- *Simplicidad:* El descriptor debería representar las características extraídas de la imagen de manera clara y sencilla para permitir una fácil interpretación de su contenido.
- *Repetibilidad:* El descriptor generado a partir de una imagen debe ser independiente del momento en el que se genere.
- *Diferenciabilidad:* Dada una imagen, el descriptor generado debe poseer alto grado de discriminación respecto de otras imágenes y al mismo tiempo contener información que permita establecer una relación entre imágenes similares.
- *Invarianza:* Cuando existen deformaciones en la representación de dos imagenes, es deseable que los descriptores que las representan aporten la

robustez necesaria para poder relacionarlas aún bajo diferentes transformaciones.

- *Eficiencia:* Es deseable que los recursos consumidos para generar el descriptor sean aceptables para poder ser utilizados en aplicaciones con restricciones críticas de espacio y/o tiempo.

Existen diferentes grados de profundidad en cuanto a la representación del contenido llevada a cabo por los descriptores, dependiendo del nivel de abstracción al que se refieran. En un nivel más bajo se encuentran los descriptores visuales, que describen características tan elementales como la forma, color, textura o movimiento entre otros. Haciendo referencia a un nivel superior se encuentran descriptores más específicos que aportan información sobre los objetos y acontecimientos de la escena. Estos últimos se apoyan en los descriptores visuales para llevar a cabo la difícil tarea de realizar una descripción semántica de las imagenes. A modo de ejemplo podemos mencionar la complejidad que supone la extracción de características relacionadas con sentimientos o sensaciones, que si bien los humanos son capaces de reconocer, no resulta evidente para los descriptores semánticos dado que dichas características no se encuentran presentes en la forma, color o textura de las imagenes.

En la siguiente sección se abordará este tema con más detalle realizando una clasificación del estado del arte de los diferentes tipos de descriptores.

En la actualidad existe un enorme interés por desarrollar descriptores audiovisuales que permitan caracterizar el contenido de las imágenes de forma automatizada. El estándar MPEG-7[7] desarrollado por MPEG (Motion Picture Expert Group) reune una colección de descriptores visuales aplicables para su implementación en tareas de recuperación de contenido multimedia, comparación y clasificación de imagenes o realización de resúmenes de vídeo.

Al mismo tiempo, coexisten otros muchos descriptores de imagen que han sido ampliamente utilizados para diferentes tareas de tratamiento de imagen y vídeo, y que si bien no pertenecen a ningún estándar, suponen una gran contribución al desarrollo de nuevas técnicas y nuevos descriptores.

[7] http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm#E9E1

2.2. Clasificación de los Descriptores de Imagen

Como se ha comentado anteriormente existen diferentes tipos de descriptores de imagen dependiendo del nivel de abstracción de la representación. Es posible clasificarlos en dos grandes grupos:

- Descriptores de información general: engloban los descriptores también llamados de bajo nivel, que proporcionan una descripción respecto del color, formas, regiones, texturas y movimientos presentes en la imagen.
- Descriptores de información de dominio específico: también llamados descriptores semánticos, proporcionan información acerca de los objetos y eventos que constituyen la escena. Lo que hacen es utilizar los descriptores de bajo nivel para cubrir el "gap" existente entre las características visuales disponibles y las diferentes categorías semánticas [5]. Un ejemplo podría ser el reconocimiento de objetos dentro de una imagen.

A su vez y dentro de los descriptores de información general, podemos clasificar a los mismos según el nivel de aplicación sobre el que actúan, es decir, sobre que regiones de la imagen realizan las distintas operaciones para generar los resultados que componen el descriptor [6]. En la figura 2.1 se representa un ejemplo que ilustra ambas categorías:

- Descriptores Globales: Resumen el contenido de la imagen en un único vector o matriz de características. Poseen la ventaja de encapsular una gran cantidad de información de la imagen requiriendo una pequeña cantidad de datos para describirla. A pesar de su simplicidad, este tipo de descriptores han resultado ser ampliamente utilizados para diferentes tareas debido entre otras cosas a su bajo coste computacional unido a unas prestaciones relativamente buenas. Un representante de esta clase es el Histograma de Color [7], descrito a continuación.
- Descriptores Locales: Son utilizados en aquellas tareas en las que una descripción local del contenido de la imagen resulta más apropiado. Actúan sobre regiones de interés, previamente calculadas o identificadas, construyendo un vector de características de esa región que tiene en cuenta la información contenida tanto en el punto de interés como en la región adyacente al mismo o vecindario. Normalmente las regiones descritas se conocen como puntos de interés, también llamados puntos destacados o *keypoints*, sin embargo estas regiones suelen referirse a bordes o pequeñas partes de la imagen. El descriptor entonces, está constituido por la totalidad de los vectores de características calculados. A modo de ejemplo podríamos mencionar el descriptor local SIFT [8].

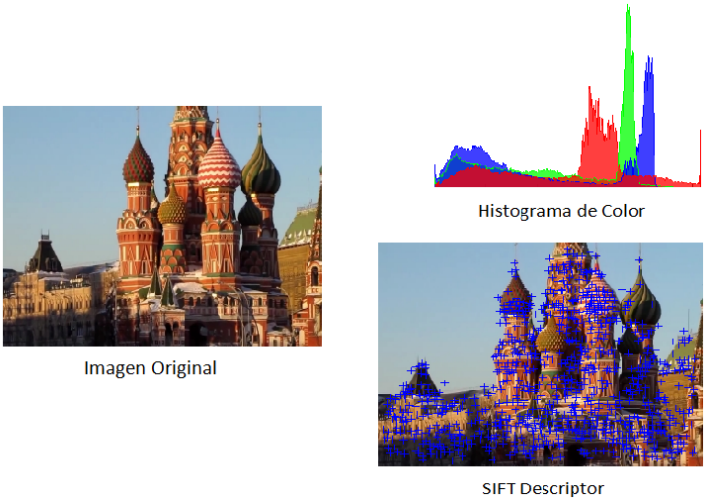

Figura 2.1: Descriptores local vs. global

Ciñéndonos al propósito de este capítulo de exponer el estado del arte sobre los descriptores visuales y del contenido específico de esta sección, nos ayudaremos de la clasificación recogida en el estándar MPEG-7 [9] y representada en la figura 2.2. Es necesario mencionar que existen diferentes clasificaciones de descriptores y, sin ánimo de ser exhaustivos y solamente a modo de ejemplo, se ha elegido la clasificación mencionada anteriormente debido a que abarca distintos tipos de descriptores exponiendo las diferentes categorías o herramientas de descripción en las que se dividen los descriptores de bajo nivel respecto de las características de la imagen sobre las que actúan. Será en el capítulo 3 donde se describirán en detalle los descriptores de imagen elegidos para este proyecto.

Herramientas para la descripción del Color (Color):

- *DominantColor Descriptor(DS).* Aunque puede ser aplicado sobre una imagen completa, su utilidad se reserva más para la representación de características locales (regiones u objetos), donde un menor número de colores son suficientes para caracterizar la región.
- *ScalableColor DS:* Representa los colores presentes en la imagen mediante un Histograma de Color HSV codificado mediante una transformación Haar. El concepto de escalabilidad se encuentra representado en la elección variable del numero de bins en los que se calcula el histograma.
- *ColorLayout DS:* Representa la distribución espacial de los colores de la imagen en el dominio frecuencial. Del mismo modo, este descriptor puede ser aplicado sobre una imagen completa o sobre regiones de interés. Presenta escalabilidad en cuanto al número de coeficientes seleccionados para la causa, si bien la recomendación apunta a 18 coeficientes de un total de 64.

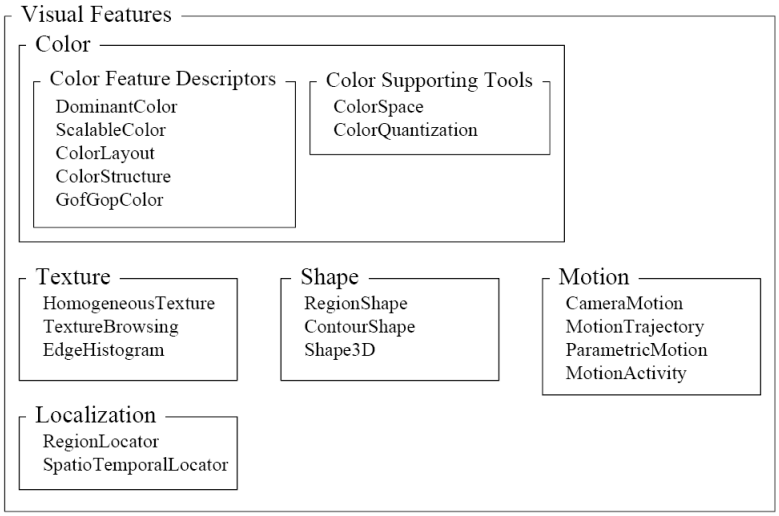

Figura 2.2: Descriptores visuales del estándar MPEG-7

- *ColorStructure DS:* Este descriptor caracteriza tanto los colores presentes como su estructura local dentro de una determinada región.
- *GoF/GoP Color DS*: Este descriptor hace uso del *ScalableColor DS* para llevar a cabo la descripción de colores presentes en un grupo de frames o imágenes de manera conjunta. El histograma final representativo del grupo de imágenes puede ser calculado mediante la media, mediana o intersección de los histogramas individuales.

Herramientas para la descripción de la Textura (Texture):

- *HomogeneousTexture DS*: Este descriptor extrae la textura presente en la imagen mediante la aplicación de diferentes filtros de Gabor sobre diferentes escalas y orientaciones quedándose con el 1º y 2º momento de la energía en el dominio frecuencial. Como resultado se obtiene un array de 62 valores codificados con 8 bits cada uno.
- *TextureBrowsing DS*: Al igual que el anterior, este descriptor hace uso de filtros de Gabor, salvo que ahora solo se seleccionan las 2 orientaciones dominantes de la imagen que son codificadas con 3 bits cada una. Al mismo tiempo se determina la regularidad (2 bits) y la aspereza (4 bits) sobre las orientaciones dominantes. Con todo ello se construye un vector de 12 bits, si bien este descriptor resulta mas adecuado para su aplicación sobre regiones que sobre la imagen completa.
- *EdgeHistogram DS*: Representa la distribución espacial de 5 tipos de bordes, 4 de ellos direccionales y el otro sin dirección, presentes en la imagen.

Herramientas para la descripción de la Forma (Shape):

- *RegionShape DS*: Este descriptor representa la forma de cualquier región de un objeto dentro de una imagen, ya sean regiones simple, como conectadas o con agujeros.
- *ContourShape DS*: Representa el contorno cerrado de una región o un objeto 2D presente en una imagen o en una secuencia de vídeo.
- *Shape3D DS*: Descripción de contornos 3D.

Herramientas para la descripción del Movimiento (Motion):

- *CameraMotion DS*: Este descriptor recoge el valor de los parámetros de los diferentes movimientos de una cámara (zoom, tilt, pan, etc) presentes en un bloque de imágenes. Cada bloque esta representado por un grupo de frames que comparten un posible movimiento de cámara.
- *MotionTrajectory DS*: Caracteriza el movimiento de un punto o región en la imagen en el dominio espacio-temporal.
- *ParametricMotion DS*: Permite caracterizar el movimiento de una región en la imagen basándose un análisis de una transformación geométrica de la evolución espacio-temporal de la región. Esta transformación se descompone en diferentes transformaciones afines como translaciones, rotaciones, zoomings, etc.
- *MotionActivity DS*: Este descriptor recoge la intensidad de la acción o el ritmo de movimiento presente en una secuencia.

Herramientas para la descripción de la Localización (Localization):

- *Region Locator DS*: Permite la localización de regiones dentro de las imágenes especificadas mediante la representación de un polígono.
- *SpatioTemporal Locator DS*: Parecido al anterior salvo que en este caso, la localización de las imágenes se extiende al dominio temporal también.

Es necesario dejar constancia de que los descriptores anteriormente descritos pertenecen al estándar MPEG-7 si bien existen muchos otros descriptores que representan o actúan en categorías semejantes. De hecho han sido propuestos muchos tipos de descriptores en la literatura y en muchos casos dependientes tanto de la aplicación como de la base de datos sobre la que se practicaban. Lo que si parece claro es que aún no se ha conseguido poner fin a la problemática de qué descriptores son mejores o peores en comparación y para que tipo de aplicaciones.

2.3. Evaluación de Descriptores

Los descriptores de imagen han sido ampliamente utilizados para llevar a cabo muy diversas tareas dentro del ámbito multimedia, como por ejemplo: reconocimiento y detección de objetos, clasificación, recuperación de imágenes basada en el contenido, resúmenes de vídeo, detección de copias, etc. Si bien los descriptores visuales han sido base fundamental para acometer dichas tareas, ofreciendo cierta capacidad de discriminación y relación, en ninguno de los casos han conseguido dar solución completa a los problemas planteados en las mismas. Han de verse por tanto, como herramientas básicas para conseguir el objetivo.

Como ya hemos mencionado en el capítulo 1, la base de este proyecto está orientada hacia la identificación de imagenes relativas a escenas cuasi-duplicadas representadas mediante diferentes transformaciones como cambios de ángulo, iluminación, etc. Es por ello que los resultados aquí obtenidos podrán ser aplicados o resultar relevantes para cualquier aplicación o estudio en el que se identifiquen necesidades similares. Como parte del estado del arte, se ha querido presentar en esta sección alguna de las evaluaciones o estudios comparativos previos sobre descriptores de imagen aplicados en diferentes contextos.

2.3.1. Evaluación de descriptores respecto de la detección de escenas similares

La problemática estudiada en este proyecto sobre la detección de escenas cuasi-duplicadas no cuenta entre la literatura con muchos exponentes. De entre los trabajos existentes cabe destacar el trabajo realizado por Mikolajczyk et al. [10] donde se expone una evaluación con respecto al comportamiento de una amplia representación de los distintos descriptores locales existentes. Se analiza la cuestión sobre la posible diferencia en el rendimiento de estos descriptores locales en base a la selección del detector de regiones de interés utilizado concluyendo resultar independiente en la mayoría de los casos. El conjunto de descriptores locales está representado por un total de 10 descriptores que implementan en total 5 detectores de regiones de interés diferentes. De entre los descriptores del estudio se destaca la creación de un nuevo representante, el descriptor GLOH (Gradient Location-Orientation Histogram), como una extensión del descriptor SIFT diferenciándose de éste en la utilización de una rejilla circular en el sistema de coordenadas polares para la creación del histograma de orientaciones de los puntos de interés y la utilización de PCA (Principal Components Analysis) [11] para la reducción de la dimensionalidad del descriptor.

El conjunto de imágenes utilizado representa 6 transformaciones geométricas y fotométricas distintas respecto de las tareas de reconocimiento de objetos y escenas: rotación y escalado de la imagen, compresión JPEG, difuminación de la imagen, cambios de iluminación y del punto de vista en las imágenes. Los resultados, que son obtenidos mediante el criterio de comparación *Precision-Recall,* presentan al descriptor GLOH como el mejor en relación al rendimiento mostrado en en la mayoría de las situaciones de análisis seguido de cerca por el

descriptor SIFT. Debido al alto coste computacional que conllevan estos descriptores, se propone como alternativa, en los casos donde el coste computacional sea una restricción a tener en cuenta o un problema, los filtros de orientación y los momentos invariantes descritos en el propio documento.

Como diferencias más significativas respecto de este proyecto se mencionan en primer lugar la variedad en la elección de los descriptores de imagen utilizados, representado por descriptores tanto locales como globales. Además en la realización de este proyecto se lleva a cabo un estudio del rendimiento de las diferentes combinaciones de los descriptores utilizados, mejorando así el rendimiento. Por último, y también de gran importancia, cabe destacar la morfología y composición de la base de datos de imágenes utilizada en el proyecto, la cual consta de una mayor cantidad y variedad de escenas e imágenes, con un menor número de transformaciones representadas.

2.3.2. Evaluación de descriptores en sistemas de recuperación basados en el contenido (CBIR)

Los sistemas CBIR, habitualmente conocidos por sus siglas de la expresión anglosajona *Content-Based Image Retrieval Systems,* abordan la problemática referente a la recuperación de la información multimedia y la búsqueda de contenidos relacionados dentro de grandes colecciones de datos. Esta aplicación muestra diversas coincidencias y puntos en común con las características de la detección de imágenes cuasi-duplicadas, objeto de este proyecto. Uno de los objetivos de la recuperación por contenido es la creación de algoritmos que sean capaces de reconocer automáticamente las características más importantes contenidas en una imagen sin intervención humana a lo largo de todo el proceso. La tarea de la recuperación basada en contenido se centra en el reconocimiento y la descripción del color, la textura, la forma, la localización espacial, las regiones de interés, y ya específicamente para imagen en movimiento se aborda la segmentación de vídeo, la extracción de fotogramas representativos o la detección de objetos específicos y de sonidos clave en el audio.

En la literatura se pueden encontrar diferentes trabajos que detallan el comportamiento de las distintas características de imagen respecto de la recuperación de imágenes. Deselaers et al. [12] presenta una evaluación cuantitativa de algunas de estas características y analiza la correlación existente entre las mismas con el objetivo tan sólo de identificar las posibles complementariedades existentes sin llevar a cabo, a diferencia del presente proyecto, ningún método de combinación de las mismas. En concreto, se evalúa el comportamiento de diferentes histogramas de color, texturas, histogramas respecto de características locales, características locales basadas en regiones y características invariantes como la amplitud espectral de la transformada de Fourier. Diferentes medidas de distancia entre las imágenes son utilizadas de acuerdo a la característica de imagen analizada, presentando a la "divergencia de Jeffrey" [13] como la distancia más recurrida para la mayoría de las características evaluadas. La comparación del rendimiento de las distintas características de imagen analizadas se lleva a

cabo mediante el uso de tasas de error (ER).

Argumentando la necesidad de seleccionar las características de imagen apropiadas dependiendo de la naturaleza del conjunto de imágenes utilizado se presenta un estudio de la correlación entre las diferentes características plasmando gráficamente los resultados obtenidos. Las conclusiones aportadas en este trabajo relatan una dependencia existente entre el rendimiento de cada una de las características analizadas y la naturaleza concreta de la base de datos de imágenes del sistema CBIR. Teniendo esto en cuenta, los histogramas de características invariantes aportan los mejores resultados en cuanto a conjuntos de imágenes en color, mientras que las características locales se comportan mejor para en el caso de conjuntos de imágenes con menor diferenciación del color como es el caso de imágenes médicas utilizadas en su evaluación.

El trabajo referenciado anteriormente [12] se ve ampliado mediante la utilización de nuevos y más grandes conjuntos de imágenes en [4], donde además el número de características analizadas se ve aumentado por la inclusión de características de forma de los objetos de las imágenes. El criterio de evaluación *ER* anterior es también sustituido por el criterio más ampliamente utilizado en tareas relativas al análisis de los sistemas CBIR de *Precision-Recall.* Es en esta nueva ampliación donde se detalla de una manera más amplia y precisa las correlaciones existentes entre las diferentes características analizadas incluyendo además una representación gráfica concisa e intuitiva de la que se pueden extraer fácilmente que características tienen propiedades similares y cuales diferentes resultando de guia para futuras combinaciones.

Siguiendo la línea del anterior trabajo, se deja constancia de la falta de una solución única respecto de las características disponibles para lidiar con las diferentes tareas relativas a los sistemas CBIR.

El histograma de color se presenta en este trabajo como una buena base sobre la que comparar el resto de las características analizadas respecto de la recuperación de imágenes en color, identificando las representaciones de características locales como aquellas que alcanzan los mejores resultados. También se destaca el hecho de que ninguna de las características basadas en la representación de la textura analizadas puede llevar a cabo una representación completa de las características de las imágenes de forma individual, si bien diferentes combinaciones de las mismas alcanzan mejores resultados.

2.3.3. Otras evaluaciones

Existen otras evaluaciones de descriptores de imagen en relación a disciplinas diferentes como son la clasificación de imágenes [14], la estimación de la posición [1] y del movimiento de los objetos [15] o el reconocimiento de objetos o clases de objetos [16] [17]. Si bien estas aplicaciones difieren en gran medida de la temática de este proyecto, se pueden establecer similitudes y encontrar resultados válidos o al menos orientativos para nuestro caso.

Respecto de la estimación del la posición, el trabajo relatado en [1] evalúa los diferentes descriptores locales elegidos en diferentes situaciones, alguna de

ellas coincidente con las transformaciones de imagen analizadas en este proyecto como son el cambio de escala (tratado en nuestro caso como variación del zoom) y cambios de la posición de la fuente de luz (en nuestro caso coincidente con los cambios de iluminación). Como resultado se muestra como el descriptor SIFT, también utilizado en este proyecto, obtiene un notable rendimiento en ambas situaciones, por delante del descriptor SURF también utilizado en nuestro caso.

3. Descriptores de Imagen Utilizados

Una vez se han descrito los distintos tipos de descriptores de imagen dependiendo tanto de las características de la imagen sobre las que actúan como del carácter de aplicación que tienen sobre la imagen, en este capítulo se exponen detalladamente los descriptores de imagen y las diferentes métricas o distancias elegidas para llevar a cabo es proyecto y cuya evaluación para la detección de escenas cuasi-duplicadas se realiza en las secciones 4.4, 4.5 y 4.6. En concreto y para cada uno de ellos se realiza una descripción conceptual donde se exponen las características que se extraen de las imágenes así como la manera en la que se lleva a cabo la representación de su contenido.

La selección está formada por los siguientes 5 descriptores: *Color Histogram, Color Layout, Color Correlogram, SIFT y SURF*. Si bien este conjunto representa cinco descriptores diferentes, en este proyecto se hace una distinción entre dos variaciones distintas del histograma de color atendiendo al sistema de color utilizado para la representación de las características de la imagen. De esta manera se presentan como *HSV histogram* y *RGB histogram* dos modalidades diferentes y que compondrán el grupo final de seis descriptores utilizados. La elección de estos descriptores frente a muchos otros así como el número que componen el estudio ha resultado ser un intento por abarcar diferentes criterios que se exponen a continuación:

- **Diferentes Grados de Complejidad**: Entre los descriptores elegidos encontramos algunos que destacan por su simplicidad como puede ser el caso del *Color Histogram*, que además ha resultado ser un sistema de referencia para comparar los resultados de otros descriptores [5], y otros que han demostrado en anteriores estudios [18] ser más complejos o costosos en lo que al coste computacional se refiere como es el caso de *SIFT Descriptor*.

- **Amplia Utilización:** La selección también ha intentado reunir a descriptores que hayan sido ampliamente utilizados en diversas tareas. Es conocida la extensa utilización de los descriptores de Color en tareas como *"content-based image retrieval"* [19, 20]. Por otro lado, descriptores más recientes como SIFT y SURF han contribuido a un gran avance en temas como el reconocimiento y detección de objetos en diversas situaciones [21, 22, 23].

- **Diversificación:** También se ha tratado de representar en esta selección, descriptores que, si bien algunos de ellos comparten las características de la imagen sobre las que actúan como puede ser el caso del Color, constituyen diferentes descriptores que representan distintas informaciones contenidas en las imágenes.

- **Perspectivas de Combinación:** Finalmente se ha tenido en cuenta la posibilidad de combinación entre los descriptores mediante un estudio

teórico previo de correlación entre descriptores [4]. Este requisito resulta indispensable para llevar a cabo una tarea importante en este proyecto como es la de buscar posibles combinaciones de los descriptores con el objetivo de mejorar los resultados obtenidos individualmente.

3.1. Histograma de Color

El primero de los descriptores seleccionados para este proyecto ha resultado ser también uno de los primeros descriptores de imagen propuestos en la literatura respecto a la característica del color implementado por primera vez por Swain *et al.* [7].

El histogram de color, en adelante histograma, representa la frecuencia de aparición de cada una de las intensidades de color presentes en la imagen, mediante la contabilidad de los pixels que comparten dichos valores de intensidad de color. El histograma está compuesto por diferentes rangos o contenedores que representan un valor o conjuntos de valores de intensidad de color.

Anterior a la etapa de contabilización de cada uno de los valores de los pixels, existe una etapa de cuantificación de los intervalos o contenedores que se refiere al proceso de reducción del número de intervalos agrupando colores cuyos valores están próximos entre si en el mismo contenedor. Esta etapa es importante en cuanto a que la cuantificación de los intervalos reduce la información representada por el descriptor sobre la imagen al mismo tiempo que reduce el tiempo de cálculo. Obviamente, cuanto mayor sea el número de intervalos, mayor poder discriminativo tendrá el descriptor. Sin embargo, un gran número de intervalos no sólo incrementará el coste computacional asociado al descriptor, sino que también resultará inapropiado e ineficiente en cuanto a las comparaciones (e.g. demasiados intervalos resultan histogramas más sensibles al ruido).

El espacio de color se define como un modelo de representación del color con respecto a los valores de intensidad. La dimensionalidad del espacio de color puede estar comprendida entre una hasta cuatro dimensiones, siendo los espacios más representativos y utilizados los formados por tres componentes o canales de color.

En el caso de este proyecto, se utilizan dos espacios de color, que a su vez resultan ser los más utilizados para este tipo de tareas: RGB (Red, Green, Blue) y HSV (Hue, Saturation y Value) formado por las componentes Hue, Saturación y Value, y que se muestran en la Figura 3.1.

El sistema RGB está formado por los colores primarios Rojo, Verde y Azul con valores entre $[0, 1]$, y cuya mezcla proporcionada resulta en el color deseado. El sistema RGB utiliza las coordenadas cartesianas como se muestra en la Figura 3.1, teniendo en consideración que la diagonal formada por los vértices $(0, 0, 0)$ negro y $(1, 1, 1)$ blanco, representa la escala de grises.

Respecto del sistema HSV, la componente Value representa la intensidad del color o brillo, la componente Hue representa lo que se conoce como tonalidad, y la componente de saturación que representa de alguna manera la densidad

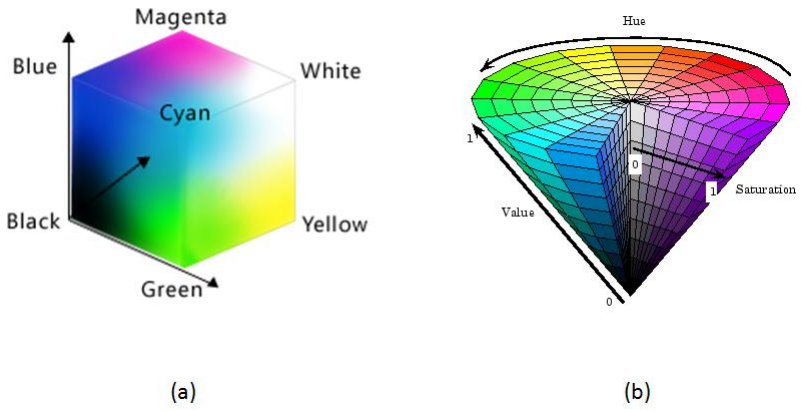

Figura 3.1: Espacios de color: (a) RGB y (b) HSV

dentro del propio color o la pureza. La resolución de las distintas componentes no es uniforme sino que se utiliza un mayor número de bits para la representación de la componente hue, que para las dos restantes, siendo suficiente dos bits en el caso de Value. Sin embargo, todas las componentes varían en un rango también normalizado de entre $[0, 1]$. El espacio de color HSV guarda una mayor relación o esta más próximo a la manera que tienen las personas de percibir el color que el espacio RGB.

La representación del histograma en uno u otro sistema conlleva ciertas restricciones que han de tenerse en cuenta en las implementaciones. Sin embargo eso no imposibilita la conversión entre ambos espacios de color. A continuación se detallan las ecuaciones necesarias para la conversión entre ambos espacios de color:

- Conversión RGB $\longrightarrow$ HSV:

$$H = \cos^{-1}\left\{\frac{\frac{1}{2}\left[(R-G)+(R-B)\right]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right\} \tag{3.1}$$

$$S = 1 - \frac{3}{R+G+B}\left[\text{mín}\,(R,G,B)\right] \tag{3.2}$$

$$V = \frac{1}{3}(R+G+B) \tag{3.3}$$

- Conversión HSV$\longrightarrow$ RGB:
  - Sector Red-Green: $(0° < H \leq 120°)$

$$B = \frac{1}{3}(1-S) \quad R = \tfrac{1}{3}\left[1+\tfrac{S\cos(H)}{\cos(60°\text{-H})}\right] \quad G = 1-(R+B) \tag{3.4}$$

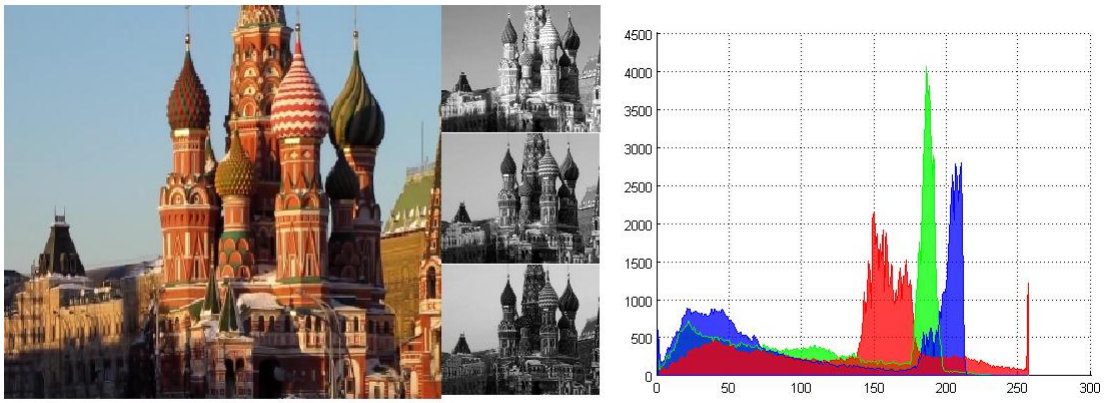

Figura 3.2: Representación del histograma RGB

a) Representación de la imagen original. b) Representación vertical de las tres componentes de color R,G,B de la imagen. c) Representación conjunta del histograma de color RGB de la imagen.

- Sector Gree-Blue: $(120° < H \leq 240°)$

$$R = \frac{1}{3}(1-S) \quad G = \tfrac{1}{3}\left[1 + \frac{S\cos(H)}{\cos(60°\text{-H})}\right] \quad B = 1-(R+B) \tag{3.5}$$

- Sector Blue-Red: $(240° < H \leq 360°)$

$$G = \frac{1}{3}(1-S) \quad B = \tfrac{1}{3}\left[1 + \frac{S\cos(H)}{\cos(60°\text{-H})}\right] \quad R = 1-(R+B) \tag{3.6}$$

Finalmente se muestra en las Figuras 3.2 y 3.3 una representación de las componentes de color e histogramas en los espacios de color RGB y HSV respectivamente.

Teniendo en cuenta que la descripción del color expuesta está formada por tres componentes, el histograma de color de una imagen, como descriptor, estará formado por la composición de los distintos histogramas de cada uno de los canales o componentes de color, construyendo así un único vector.

Comparación entre histogramas: métricas utilizadas

Una vez los histogramas de dos imágenes han sido calculados, se lleva a cabo un proceso de comparación de los descriptores con el objetivo de medir el grado de similitud que existe entre ambas. Para ello se hace uso de alguna entre las muchas métricas disponibles en la literatura para la comparación de

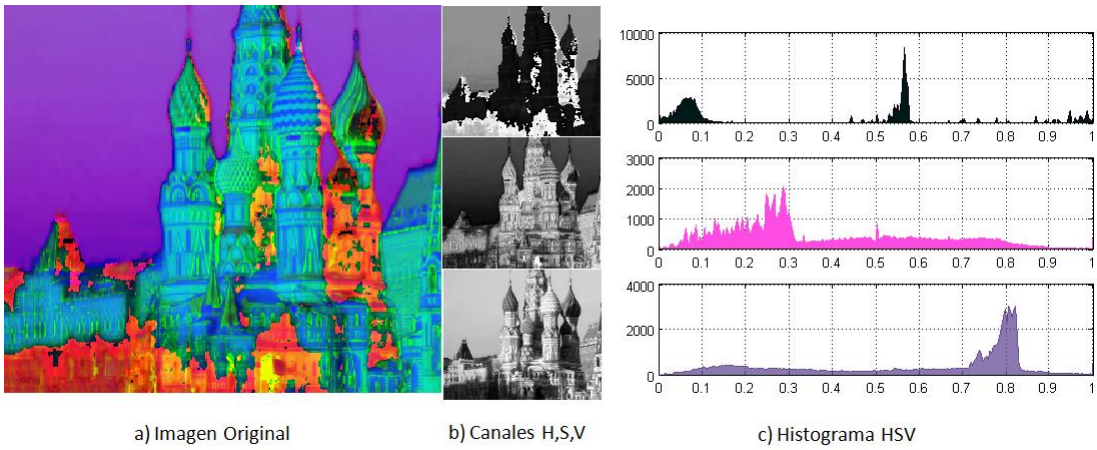

Figura 3.3: Representación del histograma HSV

a) Representación de la imagen original en el espacio de color HSV. b) Representación vertical de las tres componentes de color H,S,V de la imagen. c) Representación del histograma de cada una de las componentes de la imagen.

histogramas, y que podemos encontrar en [20] detalladas de una manera más extensa. Debido a la gran variedad, y al hecho de que para este descriptor en concreto se han utilizado diversas métricas que se expondrán con detalle en la sección 4.4, en este apartado sólo se mencionan las distintas categorías.

- Comparación de intervalos similares: Las métricas de esta categoría comparan los intervalos del mismo índice solamente sin tener en cuenta la distancia con respecto a otros intervalos.
- Comparación inter-intervalos: Las métricas aquí representadas si tienen en cuenta las comparaciones con otros intervalos que no sean estrictamente el propio.

Evolución de los histogramas

El aspecto más atractivo y ventajoso del histograma es su simplicidad y velocidad de computación, tanto en la tarea de comparación como en la de creación del descriptor. Sin embargo existen diversos inconvenientes asociados al mismo como por ejemplo la falta de consideración de información espacial de las distribuciones de color. Para subsanar algunos de los inconvenientes, como el mencionado, han surgido *mejoras o evoluciones* como es el caso de los *fuzzy color histograms* [24], *histogramas invariantes* [25] basados en los gradientes de color o finalmente *correlogramas de color* [26] cuya descripción detallada se presenta más adelante.

3.2. Color Layout Descriptor (CLD)

El estándar MPEG-7 consiste en una representación estándar de la información audiovisual que posibilita la descripción del contenido multimedia. La primera versión del estándar fue aprobada por la Organización Internacional para la Estandarización ISO/IEC en el año 2001 [9] y la última versión publicada y aprobada[8] por la ISO data del año 2004.

El estándar fue creado con el objetivo principal de llevar a cabo una gestión de los contenidos audiovisuales mediante diferentes herramientas. Estas herramientas posibilitan una descripción separada de los contenidos pero que guarda relación con ésta.

Entre las diferentes herramientas con las que cuenta el estándar para describir los aspectos principales del contenido se encuentran los descriptores, tema principal de este trabajo. A su vez, el estándar está organizado en diferentes partes, de las cuales sólo resulta de interés para este proyecto la Parte 3: Visual, que hace referencia a las estructuras básicas y descriptores que cubren diferentes características visuales como: forma, color, textura, movimiento, etc.

Uno de los distintos descriptores visuales recogidos en la Parte 3 del estándar MPEG-7 es el *Color Layout Descriptor (CLD)*[27].

El descriptor Color Layout fue diseñado para capturar la distribución espacial del color en una imagen. La representación se basa en los coeficientes de la Transformada Discreta del Coseno (DCT) sobre los valores de las componentes Y, Cb y Cr de la imagen. Esta representación se caracteriza por presentar una resolución invariante respecto del tamaño de la imagen y al mismo tiempo muy compacta.

La creación del descriptor se lleva a cabo mediante un proceso que se divide en las 4 etapas siguientes:

- División de la imagen: En la primera de las etapas, la imagen original de entrada se divide mediante una rejilla en diferentes bloques o regiones.
- Selección del color más representativo: Para cada uno de los bloques de la cuadrícula se selecciona un único color como representante de cada bloque.
- Transformada DCT: Una vez se obtiene el icono de la imagen y tras efectuar una conversión del espacio de color de la imagen original al espacio de color YCbCr, se realiza el cálculo de la DCT de cada una de las tres componentes de color, obteniendo así los llamados coeficientes de la DCT en una matriz.
- Exploración en zigzag: En esta última etapa se realiza un exploración en zigzag de los coeficientes de la matriz, con el objetivo de ponderar en mayor medida aquellos relacionados con las bajas frecuencias.

[8] http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm

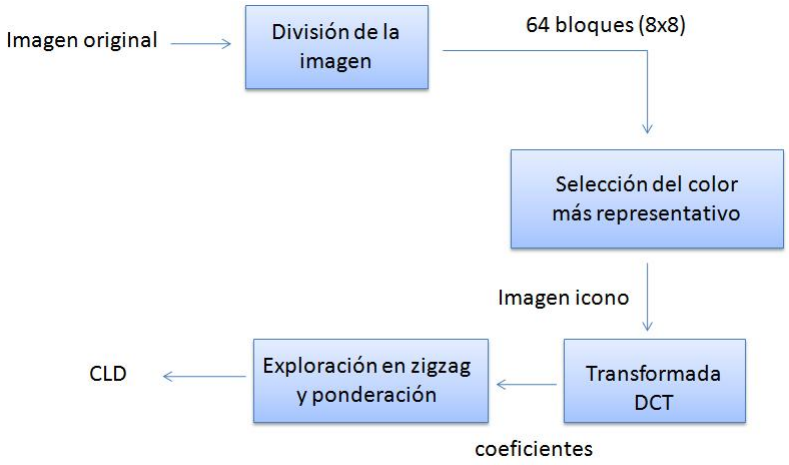

Figura 3.4: Diagrama Color Layout Descriptor

A continuación se describirán más detalladamente cada una de las etapas anteriores que componen el proceso de creación del descriptor y cuyo diagrama se muestra en la Figura 3.4.

División de la imagen

En la primera de las etapas, la imagen original es dividida en 64 (8 x 8) regiones o bloques mediante una cuadrícula cuyas dimensiones se ajustan al tamaño de la imagen. Cada uno de los bloque tiene unas dimensiones de $\left(\frac{M}{8}\ x\ \frac{N}{8}\right)$, siendo M y N las dimensiones de la imagen original. Mediante este operación se consigue la citada invarianza respecto del tamaño de la imagen original.

Figura 3.5: División de la imagen en regiones

Figura 3.6: Selección del color más representativo de cada región

Selección del color más representativo

Una vez la imagen está dividida en bloques de igual tamaño, se identifica para cada uno de ellos el color más representativo. Existen diversos métodos para calcular el color más representativo de cada bloque, siendo la media del color de los pixels comprendidos en cada bloque el recomendado por el estándar debido a su simplicidad y a que la precisión de la descripción es suficiente.

De este modo se obtiene una imagen de tamaño 8 x 8 con una apariencia borrosa, también llamada "thumbnail", y cuyo resultado puede observarse en la Figura 3.6. Esta imagen es representada por las 3 matrices de tamaño 8 x 8, donde en cada una de ellas es almacenada una componente de color de los colores representativos de cada bloque.

Transformada DCT

Como paso previo al calculo de la DCT sobre la matriz de color, se realiza una conversión del espacio de color de la imagen original al espacio de color YCbCr formado por la crominancia Y, la crominancia azul y roja Cb y Cr respectivamente.

La matriz de color es transformada mediante la aplicación de la Transformada Discreta del Coseno (DCT) obteniendo de esta manera 3 grupos de 64 coeficientes. Para calcular la DCT de una matriz 2D se utiliza la siguiente fórmula:

$$F(u,v) = \alpha_u \alpha_v \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} A_{ij} cos \frac{\pi (2i+1) u}{2M} cos \frac{\pi (2j+1) v}{2N}, \quad \begin{matrix} 0 \leq u \leq M-1 \\ 0 \leq v \leq N-1 \end{matrix}$$

$$\alpha_u = \begin{cases} \frac{1}{\sqrt{M}} & u = 0 \\ \sqrt{\frac{2}{M}} & 1 \leq u \leq M-1 \end{cases} \qquad \alpha_v = \begin{cases} \frac{1}{\sqrt{N}} & v = 0 \\ \sqrt{\frac{2}{N}} & 1 \leq v \leq N-1 \end{cases}$$

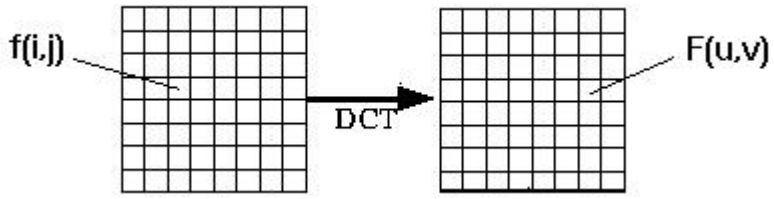

Figura 3.7: Dominio espacial y frecuencial DCT

donde (i, j)denotan las coordenadas de la matriz de color, (u, v) representan las coordenadas en el dominio transformado y A_{ij} hace referencia a la intensidad del pixel en la posición (i, j) de la matriz de entrada. En la Figura 3.7 se encuentran representados ambos dominios.

Recorrido en zigzag

En esta última etapa se realiza un recorrido en zigzag por los coeficientes de la DCT de las tres componentes YCbCr obtenidos en la etapa anterior tras haber realizado una cuantificación previa de los mismos. El motivo de este trazado tiene que ver con la ubicación de las componentes de baja frecuencia de la imagen localizadas en la parte superior izquierda de la matriz transformada lo que significa que la energía de la imagen se concentra en dicha localización.

La manera de proceder con respecto al seguimiento del trazado se puede observar en la Figura 3.8 cuyo comienzo se sitúa en la parte superior izquierda para terminal en la diagonal opuesta.

La creación del descriptor Color Layout se compone por lo tanto de 3 vectores de características que contienen los diferentes coeficientes de luminancia y crominancia, Y, Cb y Cr respectivamente, representados según el orden del recorrido en zigzag.

Según el estándar, por defecto se utilizan tan sólo los 6 primeros coeficientes relativos a la luminancia y los 3 primeros de cada una de las crominancias, aunque la inclusión de un número mayor de coeficientes está sujeta a la precisión o requisitos de la etapa de comparación.

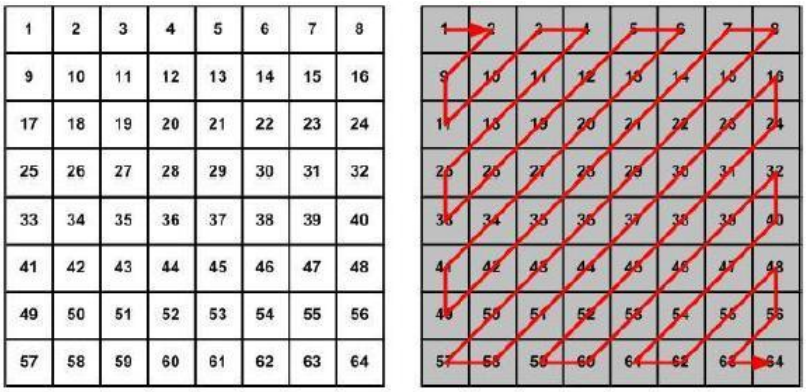

Figura 3.8: Exploración en zigzag

Comparación entre descriptores

La comparación entre descriptores tiene como objetivo evaluar el grado de similitud entre dos imagenes mediante el cálculo de la distancia entre ellos.

La función de comparación recogida en el estándar es básicamente una suma ponderada de diferencias cuadráticas entre las componentes de ambos descriptores y que es representada por la siguiente fórmula:

$$D_{CLD} = \sqrt{\sum_i \omega_i^Y \left(Y_i - Y_i'\right)^2} + \sqrt{\sum_j \omega_j^{Cb} \left(Cb_j - Cb_j'\right)^2} + \sqrt{\sum_k \omega_k^{Cr} \left(Cr_k - Cr_k'\right)^2} \tag{3.7}$$

donde (i, j, k) representan los coeficientes de las diferentes componentes (Y, Cb, Cr) respectivamente y ω_i^Y, ω_j^{Cb}, ω_k^{Cr} son los pesos elegidos para establecer las contribuciones de cada componente en la métrica. Los pesos así como el número de coeficientes puede variar dependiendo del rendimiento alcanzado en el proceso de comparación o de la importancia que se quiera dar a una u otra componente.

Observando la fórmula sobre la métrica de comparación se puede observar que dos imágenes comparten más similitudes cuanto más pequeño sea el valor de la distancia, resultando la misma imagen cuando el valor es 0.

3.3. Correlograma

El correlograma fue definido por primera vez por Huang *et al.*[26] como una nueva característica de imagen destinada a la comparación e indexación de imágenes. Esta nueva característica surgió como una alternativa más eficiente del anteriormente mencionado histograma de color.

Tras identificar las limitaciones y debilidades del histograma en cuanto a la representación de la información del color de las imágenes en tareas como la comparación de imágenes, véase las imágenes de la Figura 3.9, surgieron diversos esquemas y propuestas para mejorar los resultados obtenidos por el histograma utilizando la información espacial del color.

Algunas de las propuestas para mejorar el rendimiento del histograma han sido la de dividir la imagen en un número fijo de regiones y realizar las comparaciones mediante restricciones en cuanto a la posición relativa de las mismas (*image partitioning*). Stricker *et al.* [28] dividen la imagen en cinco regiones solapadas y realizan la extracción de los 3 momentos de color principales de cada región. De esta manera componen un vector de características representativo de cada imagen. El uso de regiones solapadas hace que los vectores de características sean relativamente insensibles ante pequeñas rotaciones o translaciones de la imagen.

Otro de los enfoques ha sido el de utilizar los histogramas con propiedades espaciales de forma local, conocida como *histogram refinement*. Pass *et al.* [29] utilizan vectores de coherencias de color (CCV) que representan la clasificación de los colores de la imagen según la coherencia de los pixels de cada color en las distintas regiones coloreadas en las que se divide la imagen.

El correlograma sin embargo, no es catalogado ni como un método de particiones ni como un esquema de refinamiento de los histogramas. Lejos de representar propiedades locales, como la posición de los pixels, o solamente propiedades globales, como la distribución del color, el correlograma tiene en cuenta tanto la correlación del color espacial de forma local junto con la distribución global de esta correlación espacial. El correlograma representa por lo tanto, el cambio de la correlación espacial de colores respecto de la distancia entre los pixels.

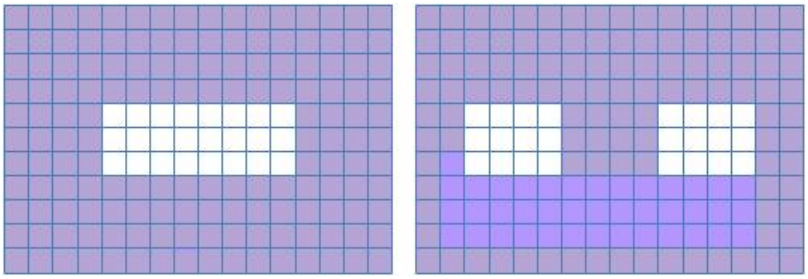

Figura 3.9: Imágenes de ejemplo correlograma

En las imágenes se puede apreciar como ambas comparten la misma distribución de colores pero sin embargo no la misma distribución espacial de los mismos. Por ello sus histogramas son iguales resultando la misma imagen, mientras que los correlogramas de ambas difieren concluyendo que ambas imágenes son diferentes como realmente son.

Notación: Sea I una imagen de dimensiones NxM cuantificada con m colores c_1, c_2, ..., c_m, para cada uno de los pixels que la componen $p = (x, y) \in I$, $p \in I_c$, lo que significa que el pixel p contiene el color c. El correlograma queda entonces definido mediante la siguiente fórmula:

$$\gamma_{c_i,c_j}^{(k)}(I) \triangleq \Pr_{p_1 \in I_{c_i}, p_2 \in I} \left[p_2 \in I_{c_j} \mid\mid p_1 - p_2 \mid = k \right] \tag{3.8}$$

donde $(i, j) \in \{1, 2, ..., m\}$, $k \in \{1, 2, ..., d\}$ y $\mid p_1 - p_2 \mid$representa la medida distancia espacial entre los pixels p_1 y p_2.

De esta manera se observa como el correlograma expresa la probabilidad de encontrar un pixel p_2 cuyo valor $I(p_2) = c_i$ a una distancia k de p_1, donde $I(p_1) = c_j$. La distancia k entre dos pixels, conocida como la distancia de cuadrícula, se determina como: $d(p_1, p_2) = \text{máx}(\mid p_{1x}, p_{2x} \mid, \mid p_{1y}, p_{2y} \mid)$.

En el ejemplo de la Figura 3.10 se muestra la cuestión principal sobre la naturaleza de este descriptor.

Comparación entre descriptores

Al igual que ocurre en el caso de otros descriptores de imagen, las métricas L_1 y L_2 han sido ampliamente utilizadas para realizar las comparaciones entre los vectores de características, alcanzando la métrica L_1[9] mejores resultados que la métrica L_2 debido a que la primera se muestra más robusta frente a valores atípicos [30]. Sin embargo Hafner *et al.* [31] introducen una medida de distancia cuadrática más sofisticada. En ella se tiene en cuenta tanto la diferencia absoluta entre ambos componentes como la diferencia relativa de esta diferencia absoluta respecto de ambas componentes. El siguiente ejemplo deja patente la mejora mencionada respecto de la norma L_1básica.

[9] Ambas métricas L_1 y L_2 pueden ser consultadas en el Anexo I de este proyecto

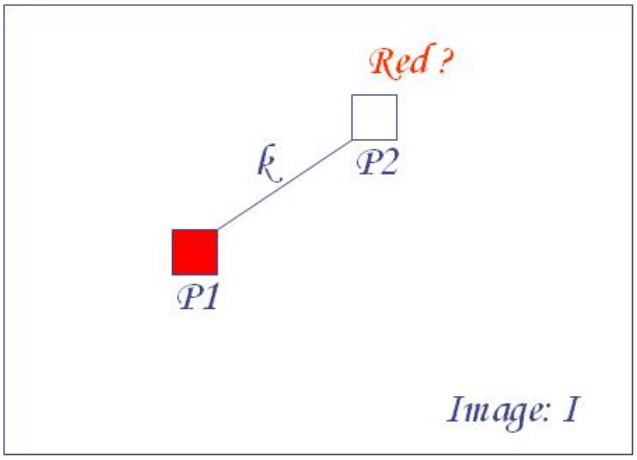

Figura 3.10: Funcionamiento del correlograma

Example. Considérese dos pares de imágenes $\langle I_1, I_2 \rangle$ y $\langle I'_1, I'_2 \rangle$.

Sean $\gamma^{(k)}_{c_1,c_2}(I_1) = 0{,}95$, $\gamma^{(k)}_{c_1,c_2}(I_2) = 0{,}9$, $\gamma^{(k)}_{c_1,c_2}(I'_1) = 0{,}25$, $\gamma^{(k)}_{c_1,c_2}(I'_2) = 0{,}2$, los correlogramas, respecto de dos colores solamente por simplicidad, de las imágenes I_1, I_2, I'_1, I'_2. Aunque la diferencia absoluta en ambos casos resulta la misma, 50, resulta más significativa con respecto al valor de ambos correlogramas para el segundo par de imágenes. Por ello la diferencia entre correlogramas debe de tener más importancia si el factor $\mid \gamma^{(k)}_{c_1,c_2}(I_1) + \gamma^{(k)}_{c_1,c_2}(I_2) \mid$ es pequeño y viceversa.

Teniendo en cuenta todo lo anterior, la distancia o métrica utilizada para comparar dos imágenes mediante sus respectivos correlogramas es implementada mediante:

$$\mid I - I' \mid_{\gamma} \triangleq \sum_{i,j \in [m], k \in [d]} \frac{\mid \gamma^{(k)}_{c_i,c_j}(I) - \gamma^{(k)}_{c_i,c_j}(I') \mid}{1 + \gamma^{(k)}_{c_i,c_j}(I) + \gamma^{(k)}_{c_i,c_j}(I')} \tag{3.9}$$

donde el factor 1 del denominador previene de posibles divisiones por 0. La inclusión de este factor obtiene una justificación teórica mediante el trabajo presentado por Haussler *et al.* [32].

La distancia entre las imágenes calculada se representa mediante un *score* o puntuación que será analizada posteriormente para la clasificación y ordenación de las imágenes comparadas en base al *score* obtenido.

3.4. Scale Invariant Feature Transform (SIFT)

El descriptor Scale Invariant Feature Transform (SIFT) fue desarrollado por Lowe [8] como un algoritmo capaz de detectar puntos característicos estables en una imagen. Estos puntos son invariantes frente a diferentes transformaciones como traslación, escala, rotación, iluminación y transformaciones afines. Originalmente fue desarrollado para el reconocimiento de objetos de manera general y realiza la correspondencia entre puntos basada en los vectores de características de cada punto que componen el descriptor de la imagen.

El algoritmo SIFT se compone principalmente de cuatro etapas que se describen siguiendo la implementación de Lowe [8]:

1. **Detección de Extremos en el Espacio Escala:** La primera etapa del algoritmo realiza una búsqueda sobre las diferentes escalas y dimensiones de la imagen identificando posibles puntos de interés, invariantes a los cambios de orientación y escalado. Esto se lleva a cabo mediante la función DoG (*Difference-of-Gaussian*).

2. **Localización de los Puntos Clave:** Para seleccionar los puntos clave, también llamados puntos de interés, de forma precisa, se aplica una medida de estabilidad sobre todos ellos para descartar aquellos que no sean adecuados.

3. **Asignación de la Orientación:** Se asignan una o más orientaciones a cada punto de interés extraído de la imagen basándose en las direcciones locales presentes en la imagen gradiente. Todas las operaciones posteriores sor realizadas sobre los datos transformados según la orientación, escala y localización dentro de la imagen asignados en esta etapa, proporcionando así la invarianza respecto de estas transformaciones.

4. **Descriptor del Punto de Interés:** La última etapa hace referencia a la representación de los puntos clave como una medida de los gradientes locales de la imagen en las en las proximidades de dichos puntos clave y respecto de una determinada escala. Cada punto de interés corresponde a un vector de características compuesto por 128 elementos, que le confiere una invarianza parcial a deformaciones de forma así como cambios de iluminación.

La estabilidad de los puntos de interés es importante debido a que la comparación realizada entre objetos pertenecientes a dos imágenes diferentes se lleva a cabo mediante la comparación de los mismos puntos de interés. Para asegurar esta estabilidad, Brown y Lowe [33] proponen una función 3D para eliminar aquellos puntos que se encuentren en bordes o que presenten bajo contraste, ya que son mas susceptibles al ruido.

Detección de extremos en el espacio escala

La primera de las etapas tiene como objetivo obtener puntos candidatos de la imagen que puedan ser identificados de forma repetida bajo diferentes

vistas del mismo objeto. El descriptor SIFT es construido a partir del espacio-escala Gaussiano de la imagen original, en el cual se pueden detectar de manera efectiva las posiciones de los puntos claves, invariantes a cambios de escala de la imagen. El espacio-escala Gaussiano de una imagen $L(x, y, \sigma)$ es definido como la convolución de funciones 2D Gaussianas $G(x, y, \sigma)$ de diferentes valores σ con la imagen original $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{3.10}$$

siendo (x, y) las coordenadas espaciales y σ el factor de escala.

El algoritmo utiliza la función DoG (Diferencia de Gaussiana) que se forma a partir de la derivada escalar de la Gaussiana escalada espacialmente. Esta función DoG $D(x, y, \sigma)$ se obtiene mediante la sustracción de escalas posteriores en cada octava:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{3.11}$$

donde k es una constante multiplicativa del factor de escala. La función DoG es utilizada por varias razones. En primer lugar porque es una función eficiente en cuanto a coste computacional se refiere: Las imagenes suavizadas $L(x, y, \sigma)$ son calculadas para la descripción de características en el espacio-escala, y por lo tanto, D puede obtenerse como una simple resta. Además, Mikolajczyk [10] asegura que los máximos y mínimos del Laplaciano de la Gaussiana respecto de una escala normalizada produce las características de imagen más estables características de imagen en comparación con otras funciones como el Gradiente, el Hessiano o el Harris Corner Detector, pudiéndose aproximar el Laplaciano de la Gaussiana de escala normalizada mediante la función DoG.

Al conjunto de las imágenes Gaussianas suavizadas junto con las imágenes DoG se le llama octava. El conjunto de las octavas es construido mediante el muestreo sucesivo de la imagen original por un factor de 2. Cada una de las octavas (i.e., duplicando σ) es a su vez dividida en un número entero de subniveles o escalas s. Una vez se ha procesado una octava completa, la primera imagen de la siguiente octava se obtiene mediante el muestreo de la primera de las imágenes de la octava predecesora con un valor de σ del doble respecto a la actual. Esto se traduce en una gran eficiencia del algoritmo para un número de escalas pequeño. El proceso descrito puede verse representado en la Figura 3.11. Es importante tener en cuenta que la imagen original es expandida en el inicio del proceso para crear más puntos de muestreo que en la imagen original, por lo que la imagen resulta duplicada en tamaño antes de construir el primer nivel de la pirámide.

Dado que el espacio-escala $L(x, y, \sigma)$ representa la misma información a diferentes niveles de escala, el modo particular del muestreo permite una reducción de la redundancia. De esta manera se producen $s + 3$ imágenes por cada una de las octavas y por lo tanto $s + 2$ DoG imágenes donde se llevará a cabo la búsqueda de extremos. De acuerdo con los resultados de Lowe, es el valor de $s = 3$ el que mejores resultados consigue, con lo que es el que se utiliza en este

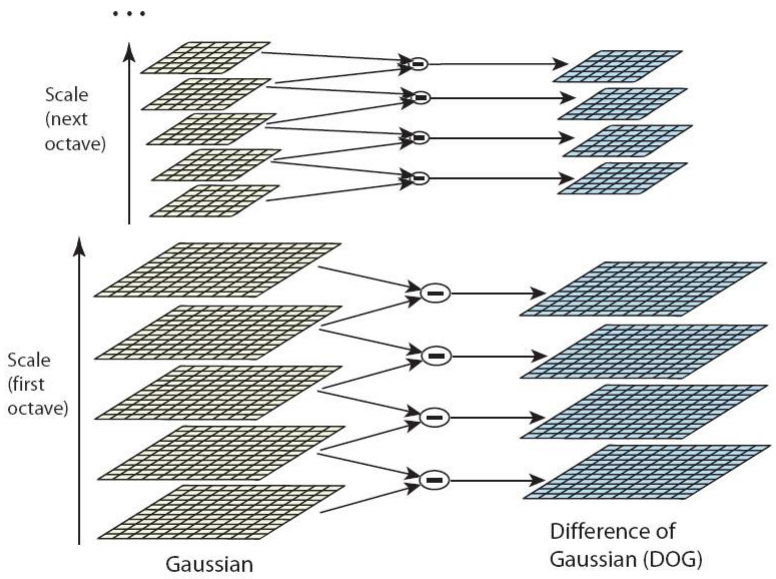

Figura 3.11: Creación del espacio-escala Gaussiano.

En cada una de las escalas, también llamadas octavas, la imagen se convoluciona repetidamente con funciones gaussianas para producir el conjunto de imagenes gaussianas mostradas en la parte izquierda de la imagen. Las imagenes obtenidas son substraídas en parejas adyacentes para producir las imagenes diferencia-de-gaussiana mostradas a la derecha. Después de cada octava, las imagenes Gaussianas son muestreadas por un factor de 2, y se repite el proceso. Fuente David Lowe [8].

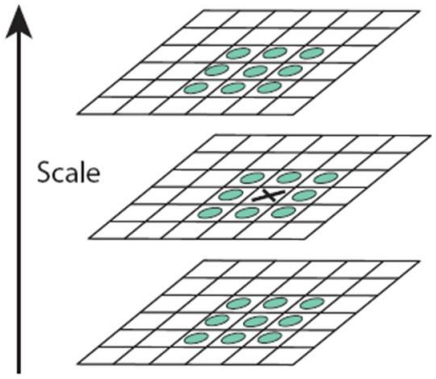

Figura 3.12: Localización de máximos y mínimos locales.

Los máximos y mínimos de las imágenes diferencia-de-gaussiana son detectados mediante la comparación de un pixel (marcado con X) con sus 26 vecinos en las regiones de 3x3 de las escalas actual y adyacentes (marcados en azul). Fuente: David Lowe [8].

proyecto. Con esto se obtienen 6 imágenes Gaussianas suavizadas y 5 imágenes DoG por cada octava. Respecto del otro parámetro por determinar, σ referente al muestreo de la escala de suavizado, se ha adoptado por seguir el mismo criterio anterior en base a los resultados de Lowe, donde se determina un valor de $\sigma = 1{,}6$. Una explicación más detallada se encuentra en [8].

Para detectar los máximos y los mínimos locales de cada punto de la imagen $D(x, y, \sigma)$ se compara el valor de éste con el de los puntos vecinos, en concreto, con el de sus 8 vecinos más próximos de la imagen D donde se encuentra el punto más los 9 vecinos de cada una de las imágenes D de nivel superior e inferior como se muestra en la Figura 3.12. Si el valor resulta ser superior o inferior al de todos sus vecinos, se identifica el punto como máximo o mínimo local respectivamente.

Localización de puntos clave estables

Una vez los puntos clave candidatos han sido calculados, en esta segunda etapa se realiza un estudio de su estabilidad. Los puntos no firmemente situados sobre los bordes o aquellos con bajo contraste son bastante vulnerables al ruido y por lo tanto no podrán ser detectados bajo pequeños cambios de iluminación o variación del punto de vista de la imagen. Para excluirlos, Lowe utiliza los

siguientes criterios:

- Para eliminar los puntos con bajo contraste, se aplica un proceso de umbralización por el cual los puntos cuyo valor sea menor que dicho umbral D serán excluidos de la siguiente etapa por no considerarse suficientemente estables. En este proyecto se utiliza el valor de $D = 0{,}03$ recomendado por Lowe.

- Los puntos situados sobre bordes de manera difusa, conllevan un alto grado de inestabilidad incluso ante pequeños ruidos. Para llevar a cabo su eliminación, se utiliza la propiedad de la función DoG atendiendo a la gran curvatura que presenta en la dirección paralela al borde y la pequeña curvatura que se observa en la dirección perpendicular. Estas respuestas tan características se pueden estudiar mediante el cálculo de la matriz del Hessiano sobre la localización y escala del punto en estudio:

$$H = \begin{bmatrix} \frac{\partial^2 D}{\partial x^2} & \frac{\partial^2 D}{\partial x \partial y} \\ \frac{\partial^2 D}{\partial x \partial y} & \frac{\partial^2 D}{\partial y^2} \end{bmatrix} \tag{3.12}$$

donde D es la imagen DoG $D(x, y, \sigma)$ respecto de la escala s. Las derivadas se calculan mediante la resta del valor de los puntos vecinos. Se puede demostrar que la siguiente desigualdad permite la localización de los puntos en los bordes:

$$\frac{\left(\frac{\partial^2 D}{\partial x^2} + \frac{\partial^2 D}{\partial y^2}\right)^2}{\left(\frac{\partial^2 D}{\partial x^2} \times \frac{\partial^2 D}{\partial y^2}\right) - \left(\frac{\partial^2 D}{\partial x \partial y}\right)^2} < \frac{(r+1)^2}{r} \tag{3.13}$$

por lo tanto aquellos puntos que no satisfagan dicha desigualdad serán descartados debido a su inestabilidad. El valor fijado es de $r = 10$ al igual que en el paper de referencia [8]. Tras descartar los puntos inestables, al resto de puntos clave se les asignará una orientación.

Asignación de la orientación

La característica principal de los puntos SIFT es que éstos son invariantes a una serie de transformaciones sobre las imágenes.

La invarianza respecto de la rotación se consigue mediante la asignación a cada uno de los puntos una orientación basada en las propiedades locales de la imagen y representando el descriptor respecto de esta orientación. Para cada uno de los puntos de interés se calcula la magnitud del gradiente, m, y su orientación, θ, mediante las siguientes ecuaciones:

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \tag{3.14}$$

$$\theta(x, y) = \arctan \frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)} \tag{3.15}$$

donde L representa la imagen gaussiana suavizada cuya escala resulta más próxima a la escala del punto de interés actual.

Respecto de la orientación del gradiente, se crea un histograma con 36 bins, cada uno de ellos con una longitud de 10° para cubrir el rango de los 360° posibles. El *bin* cuyo valor es mas alto se corresponde con la dirección dominante del gradiente y por lo tanto es elegido como orientación dominante. Sin embargo se ha de tener en cuenta la posibilidad de que exista más de una dirección dominante. Es por ello que cualquier *bin* con un valor de más del 80 % del valor de la magnitud principal se considerará también como dirección dominante. Los puntos que contengan más de una dirección dominante supondrán una mayor estabilidad al mismo. Para una mayor precisión se utiliza una parábola para mediante la interpolación de los 3 valores más altos del histograma obtener el valor del pico.

Las orientaciones principales del histograma se asignan al punto de interés para que así el descriptor quede representado respecto de éstos.

Descriptor del punto de interés

Las etapas anteriores han dotado a los puntos de interés seleccionados de invarianza respecto de la orientación, escalada y localización respecto de la imagen. En esta última etapa se crea un vector de características para cada uno de los puntos de interés que contiene una estadística local de las orientaciones del gradiente de la escala de espacio gaussiano. Se realiza un muestreo de las orientaciones y magnitudes del gradiente de la imagen sobre regiones de 16×16 alrededor del punto de interés. Este proceso es similar al de la etapa anterior, donde ahora cada una de las muestras son ponderadas tanto por la magnitud de su gradiente como por una función 3D gaussiana evitando así cambios bruscos en el descriptor ante pequeños cambios en la posición de la ventana y al mismo tiempo asignando menor énfasis a los puntos más alejados del punto de interés.

El valor σ de la función gaussiana se fija como 1,5 veces el tamaño de la región de cálculo para el punto de interés.

Se analizan las muestras de cada región de 16×16 formando histogramas de orientaciones resumiendo el contenido en sub-regiones de 4×4 como se puede ver en la Figura 3.13. Cada uno de los histogramas se compone de 8 bins, que almacenan las orientaciones posibles proporcionales a 45° donde la magnitud de cada flecha representa el valor acumulado para cada *bin*. Por lo tanto se obtienen 16 histogramas respecto de las orientaciones de los puntos de cada región para cada uno de los puntos de interés.

Finalmente el descriptor de cada punto de interés está formado por un vector que contiene los valores de las 8 orientaciones de los 4×4 histogramas componiendo un vector de características de $4 \times 4 \times 8 = 128$ elementos.

De manera añadida, el vector de características es modificado para dotarlo de cierta robustez frente a cambios de iluminación. Los cambios de iluminación afectan en mayor medida a la magnitud del gradiente y no a la orientación, por

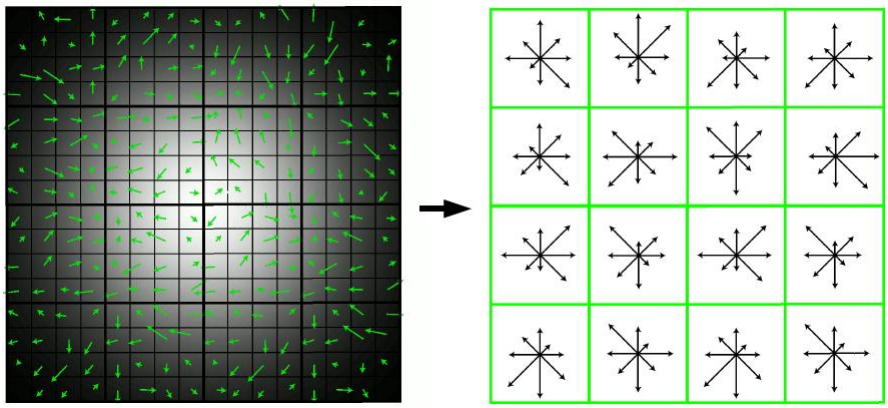

Figura 3.13: Descriptor de los puntos de interés

lo que se busca una representación de esta magnitud que minimice estos efectos. Para ello se lleva a cabo un proceso de normalización en los que ahora los cambios de contraste (pixels multiplicados por una constante) quedan neutralizados, mientras que los cambios en la luminosidad (suma de una constante con los pixels) no afecta a los valores del gradiente que se calcula como diferencias entre pixels. Si bien esta normalización no confiere invarianza respecto de los cambios de iluminación, si se consigue paliar los efectos que estos producen.

Finalmente se limita el valor de cada componente de magnitud de gradiente a un valor máximo para que tenga un mayor peso la orientación frente a la magnitud del gradiente. Siguiendo los parámetros de Lowe [8], el valor del umbral es de 0,2. Luego se vuelve a normalizar de nuevo a una amplitud unidad.

Correspondencia entre puntos clave (*matching*)

El término *matching* entre imágenes tiene como finalidad el cálculo de un valor que represente el grado de similitud entre las dos imágenes, y que a continuación se puedan establecer las diferentes conclusiones. El cálculo de este valor, representado como distancia y conocido también como *score,* se realiza mediante la aplicación de una métrica o fórmula de la distancia entre ambas imágenes. Previo paso del cálculo del *score*, es necesario establecer las correspondencias entre los puntos clave.

La correspondencia entre puntos clave se lleva a cabo mediante el cálculo de la distancia euclídea entre los vectores de características pertenecientes a diferentes puntos de interés. Este cálculo genera a su vez otro valor que será utilizado para determinar cual de los puntos de la imagen comparada se corresponde con su homólogo, en el caso de existir, de la primera de las imágenes.

Supongamos que queremos realizar el *matching* de puntos entre dos imágenes I_1 e I_2. Para cada uno de los puntos clave pertenecientes a I_1, se seleccionan los dos mejores candidatos de entre todos los puntos clave pertenecientes a I_2

Figura 3.14: Diagrama de bloques del descriptor SIFT

Figura 3.15: Representación del matching para el descriptor SIFT

mediante el criterio de máxima similitud. Este criterio establece que los mejores candidatos para realizar el matching con el punto clave p_1 perteneciente a I_1 cuyo vector de características es v_1, son los puntos clave p'_1 y p'_2 pertenecientes a I_2 cuyos vectores de características v'_1 y v'_2 representan las distancias euclídeas mínimas d_1 y d_2 respectivamente respecto de v_1. Si la relación d_1/d_2 entre las distancias mencionadas es suficientemente pequeña, entonces se establece el matching entre los puntos p_1 y p'_1 pertenecientes a cada una de las imágenes. De acuerdo con [8], se establece un umbral de 0.76 para el ratio d_1/d_2.

Esta estrategia de matching recibe el nombre de *"el vecino más próximo"*. Finalmente la puntuación o *score* entre las dos imágenes se obtiene mediante una relación que tiene en cuenta el número total de puntos *matcheados* entre ambas imágenes.

Mediante el diagrama de bloques de la Figura 3.14 se representan las etapas de funcionamiento del proceso de comparación entre dos imágenes y el cálculo del *score* o puntuación entre las mismas.

Finalmente se puede observar en la Figura 3.15 el resultado del matching de puntos clave entre dos imágenes para el descriptor SIFT.

Los resultados de este proceso serán presentados en detalle en la sección 4.4.

3.5. Speeded Up Robust Features (SURF)

El descriptor SURF, cuyo acrónimo hace referencia al título, Speeded-Up Robust Features, fue desarrollado por Herbert Bay et al. [34] como un detector de puntos de interés y descriptor robustos. El descriptor SURF guarda cierta similitud con la filosofía del descriptor SIFT [8], si bien presenta notables diferencias que quedarán patentes con la siguiente exposición sobre su desarrollo. Los autores afirman sin embargo que este detector y descriptor presentan principalmente 2 mejoras resumidas en los siguientes conceptos:

- Velocidad de cálculo considerablemente superior sin ocasionar perdida del rendimiento.
- Mayor robustez ante posibles transformaciones de la imagen.

Estas mejoras se consiguen mediante la reducción de la dimensionalidad y complejidad en el cálculo de los vectores de características de los puntos de interés obtenidos, mientras continúan siendo suficientemente característicos e igualmente repetitivos.

A continuación se describirán en detalle las etapas para la creación de los descriptores SURF, si bien antes se presentan a modo de resumen previo las diferencias más originales respecto del descriptor SIFT:

- La normalización o longitud de los vectores de características de los puntos de interés es considerablemente menor, concretamente se trata de vectores con una dimensionalidad de 64, lo que supone una reducción de la mitad de la longitud del descriptor SIFT.
- El descriptor SURF utiliza siempre la misma imagen, la original.
- Utiliza el determinante de la matriz Hessiana para calcular tanto la posición como la escala de los puntos de interés

Detección de puntos de interés

La primera de las etapas del descriptor SURF es análoga a la del descriptor SIFT en cuanto a la detección de puntos de interés se refiere, si bien el procedimiento para su obtención se basa en diferencias sustanciales que se detallan a continuación.

El descriptor SURF hace uso de la matriz Hessiana, más concretamente, del valor del determinante de la matriz, para la localización y la escala de los puntos. El motivo para la utilización de la matriz Hessiana es respaldado por su rendimiento en cuanto a la velocidad de cálculo y a la precisión. Lo realmente novedoso del detector incluido en el descriptor SURF respecto de otros detectores es que no utiliza diferentes medidas para el cálculo de la posición y la escala de los puntos de interés individualmente, sino que utiliza el valor del determinante de la matriz Hessiana en ambos casos. Por lo tanto dado un punto

$p = (x, y)$ de la imagen I , la matriz Hessiana $H(p, \sigma)$ del punto p perteneciente a la escala σ se define como:

$$H(p, \sigma) = \begin{bmatrix} L_{xx}(p, \sigma) & L_{xy}(p, \sigma) \\ L_{xy}(p, \sigma) & L_{yy}(p, \sigma) \end{bmatrix} \tag{3.16}$$

donde $L_{xx}(p, \sigma)$ representa la convolución de la derivada parcial de segundo orden de la Gaussiana $\frac{\partial^2}{\partial x^2} g(\sigma)$ con la imagen I en el punto p. De manera análoga ocurre con los términos $L_{xy}(p, \sigma)$, $L_{yy}(p, \sigma)$ de la matriz.

A pesar de que los filtros gaussianos son óptimos para el análisis del espacio-escala [35], se ha implementado una alternativa a los filtros gaussianos en el detector SURF debido a una serie de limitaciones de estos filtros (como la necesidad de ser discretizados, la falta de prevención total del indeseado efecto aliasing, etc.): los filtros tipo caja (de sus siglas en inglés box-filters)[36]. Estos nuevos filtros aproximan las derivadas parciales de segundo orden de las gaussianas y pueden ser evaluados de manera muy rápida usando imágenes integrales, independientemente del tamaño de éstas . Las imágenes integrales, cuya definición se encuentra ampliamente detallada en [37, 38], son calculadas mediante la siguiente fórmula:

$$Ii_{\sum}(x, y) = \sum_{i=1}^{i \leq x} \sum_{j=1}^{j \leq y} I(i, j) \tag{3.17}$$

donde (x, y) representan la posición del punto en la imagen y $Ii(x, y)$representa la intensidad de la imagen en el punto.

Una vez la imagen integral ha sido creada, se puede calcular la suma de las intensidades de una región mediante una simple operación, como se puede observar en la Figura 3.16:

$$\sum I = Ii_D + Ii_A + Ii_B + Ii_C \tag{3.18}$$

De esta forma, el tiempo necesario para el calculo de las operaciones de convolución es independiente del tamaño de la imagen.

El espacio escala del descriptor SIFT descrito anteriormente, se crea a partir de imágenes suavizadas repetidamente mediante la aplicación de un filtro gaussiano y que posteriormente se submuestrean para alcanzar un nivel más alto dentro de la pirámide de dicho espacio escala. Sin embargo en el caso del detector SURF, debido a la utilización de filtros de tipo caja e imágenes integrales, no es necesario aplicar el mismo filtro iterativamente a la salida de una capa filtrada previamente, sino que se pueden aplicar dichos filtros de cualquier tamaño a la misma velocidad directamente sobre la imagen original. De este modo resulta que el espacio escala es analizado mediante la elevación del tamaño del filtro, en vez de reducir el tamaño de la imagen como es el caso del detector SIFT como se puede observar en la Figura 3.17.

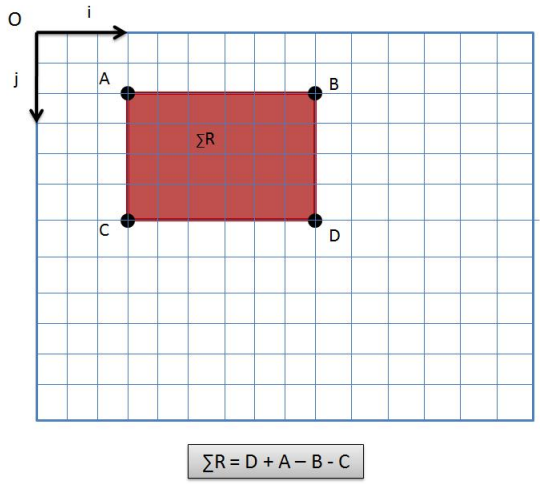

Figura 3.16: Representación de la intensidad de una región respecto de la imagen integral

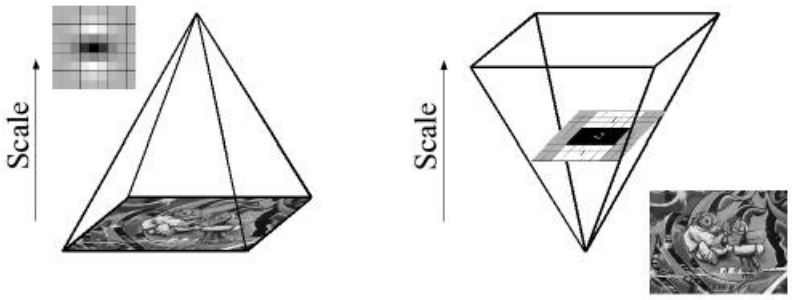

Figura 3.17: Espacio escala SIFT vs. SURF

Las aproximaciones de las derivadas parciales se denotan como D_{xx}, D_{xy}, y D_{yy}. En cuanto al determinante de la matriz Hessiana, éste queda definido de la siguiente manera:

$$det\left(H_{aprox.}\right) = D_{xx}D_{yy} - \left(0{,}9D_{xy}\right)^2 \tag{3.19}$$

donde el valor de 0,9 está relacionado con la aproximación del filtro gaussiano.

En la Figura 3.18 se puede observar la representación de la derivada parcial de segundo orden de un filtro gaussiano discretizado y la aproximación de la derivada implementada en el caso del descriptor SURF.

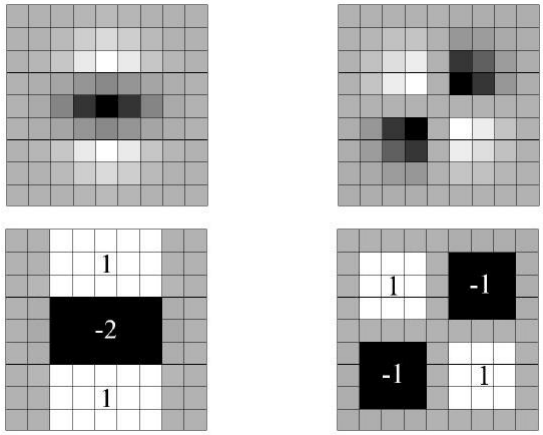

Figura 3.18: Derivadas parciales de segundo orden de un filtro gaussiano y su aproximación

Derivada parcial de segundo orden discretizadas en la dirección xy e y, L_{xy} y L_{yy} respectivamente (arriba), y sus respectivas aproximaciones D_{xy} y D_{yy} (abajo).

La imagen de salida obtenida tras la convolución del la imagen original con un filtro de dimensiones 9 x 9, que corresponde a la derivada parcial de segundo orden de una gaussiana con σ = 1,2, es considerada como la escala inicial o también como la máxima resolución espacial (s = 1,2, correspondiente a una gaussiana con σ = 1,2). Las capas sucesivas se obtienen mediante la aplicación gradual de filtros de mayores dimensiones, evitando así los efectos de aliasing en la imagen.

El espacio escala para el descriptor SURF, al igual que en el caso del descriptor SIFT, está divido en octavas. Sin embargo, en el descriptor SURF, las octavas están compuestas por un número fijo de imágenes como resultado de la convolución de la misma imagen original con una serie de filtros cada más grandes. El incremento o paso de los filtros dentro de una misma octava es el doble respecto del paso de la octava anterior, al mismo tiempo que el primero de

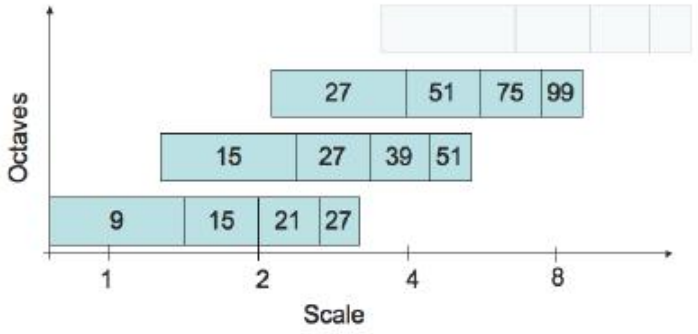

Figura 3.19: Representación gráfica de la longitud de los filtros de diferentes octavas

los filtros de cada octava es el segundo de la octava predecesora. De esta manera obtenemos las siguientes series de octavas con sus respectivos filtros como se muestra en la Figura 3.19:

- Octava inicial: $9x9 \xrightarrow{6} 15x15 \xrightarrow{6} 21x21 \xrightarrow{6} 27x27$
- Octava siguiente: $15x15 \xrightarrow{12} 27x27 \xrightarrow{12} 39x39 \xrightarrow{12} 51x51$
- Octava siguiente: $27x27 \xrightarrow{24} 51x51 \xrightarrow{24} 75x75 \xrightarrow{24} 99x99$
- Y así sucesivamente...

Finalmente para calcular la localización de todos los puntos de interés en todas las escalas, se procede mediante la eliminación de los puntos que no cumplan la condición de máximo en un vecindario de 3 x 3 x 3. De esta manera, el máximo determinante de la matriz Hessiana es interpolado en la escala y posición de la imagen. En este punto se da por concluida la etapa de detección de los puntos de interés.

Asignación de la orientación

La siguiente etapa en la creación del descriptor corresponde a la asignación de la orientación de cada uno de los puntos de interés obtenidos en la etapa anterior. Es en esta etapa donde se otorga al descriptor de cada punto la invarianza ante la rotación mediante la orientación del mismo.

El primer paso para otorgar la mencionada orientación consiste en el cálculo de la respuesta de Haar en ambas direcciones x e y mediante las funciones representadas en la Figura 3.20.

El área de interés para el cálculo es el área circular centrada en el punto de interés y de radio $6s$, siendo s la escala en la que el punto de interés ha sido

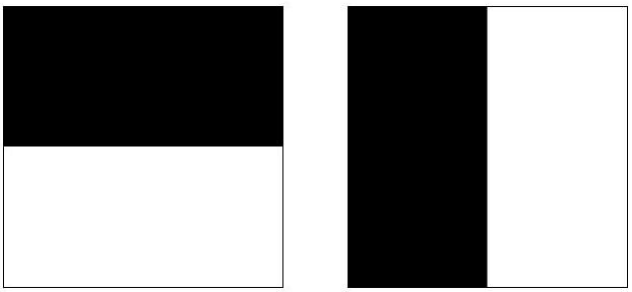

Figura 3.20: Filtros de Haar empleados en el descriptor SURF

Funciones de Haar para el cálculo de las respuestas en la dirección x (izquierda) e y (derecha). El color negro identifica el valor -1 y el color blanco el valor +1.

detectado. De la misma manera, la etapa de muestreo depende de la escala y se toma como valor s. Respecto de las funciones onduladas de Haar, se toma el valor $4s$, por tanto dependiente también de la escala, como referencia, donde a mayor valor de escala mayor es la dimensión de las funciones onduladas.

Tras haber realizado todos estos cálculos, se utilizan imágenes integrales nuevamente para proceder al filtrado mediante las máscaras de Haar y obtener así las respuestas en ambas direcciones. Son necesarias únicamente 6 operaciones para obtener la respuesta en la dirección x e y. Una vez que las respuestas onduladas han sido calculadas, son ponderadas por una gaussiana de valor $\sigma = 2{,}5s$ centrada en el punto de interés. Las respuestas son representadas como vectores en el espacio colocando la respuesta horizontal y vertical en el eje de abscisas y ordenadas respectivamente. Finalmente, se obtiene una orientación dominante por cada sector mediante la suma de todas las respuestas dentro de una ventana de orientación móvil cubriendo un ángulo de $\frac{\pi}{3}$ siguiendo las especificaciones recomendadas por el autor [34]. La representación de este puede observarse en la Figura 3.21.

La orientación final del punto de interés será finalmente aquella cuyo vector sea el más grande dentro de los 6 sectores en los que han sido dividida el área circular alrededor del punto de interés.

Creación del descriptor

Es en esta última etapa del proceso donde se concreta la creación del descriptor SURF.

Se construye como primer paso una región cuadrada de tamaño $20s$ alrededor del punto de interés y orientada en relación a la orientación calculada en la etapa anterior. Esta región es a su vez dividida en 4 x 4 sub-regiones dentro de cada una de las cuales se calculan las respuestas de Haar de puntos con una separación de muestreo de 5 x 5 en ambas direcciones. Por simplicidad, se

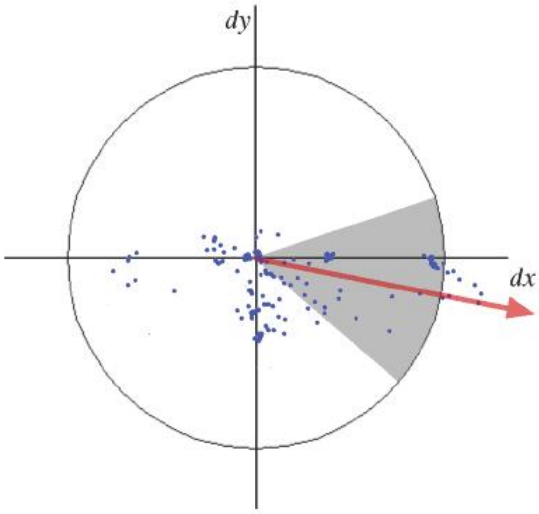

Figura 3.21: Asignación de la orientación de cada sector
Un vector de orientación es calculado como representante de todas las respuestas gaussianas de Haar en los puntos de muestreo contenidos en cada sector circular de valor $\frac{\pi}{3}$.

consideran d_x y d_y las respuestas de Haar en las direcciones horizontal y vertical respectivamente relativas a la orientación del punto de interés. En la Figura 3.22 están representadas tanto las respuestas de Haar en cada una de las sub-regiones como las componentes d_x y d_y uno de los vectores.

Para dotar a las respuestas d_x y d_y de una mayor robustez ante deformaciones geométricas y errores de posición, éstas son ponderadas por una gaussiana de valor $\sigma = 3{,}3s$ centrada en el punto de interés.

En cada una de las sub-regiones se suman las respuestas d_x y d_y obteniendo así un valor de d_x y d_y representativo por cada una de las sub-regiones. Al mismo tiempo se realiza la suma de los valores absolutos de las respuestas $\mid d_x \mid$ y $\mid d_y \mid$ en cada una de las sub-regiones, obteniendo de esta manera, información de la polaridad sobre los cambios de intensidad.

En resumen, cada una de las sub-regiones queda representada por un vector ν de componentes:

$$\nu = \left(\sum d_x, \sum d_y, \sum \mid d_x \mid, \sum \mid d_y \mid \right) \tag{3.20}$$

y por lo tanto, englobando las 4 x 4 sub-regiones, resulta un descriptor SURF con una longitud de 64 valores para cada uno de los puntos de interés identificados.

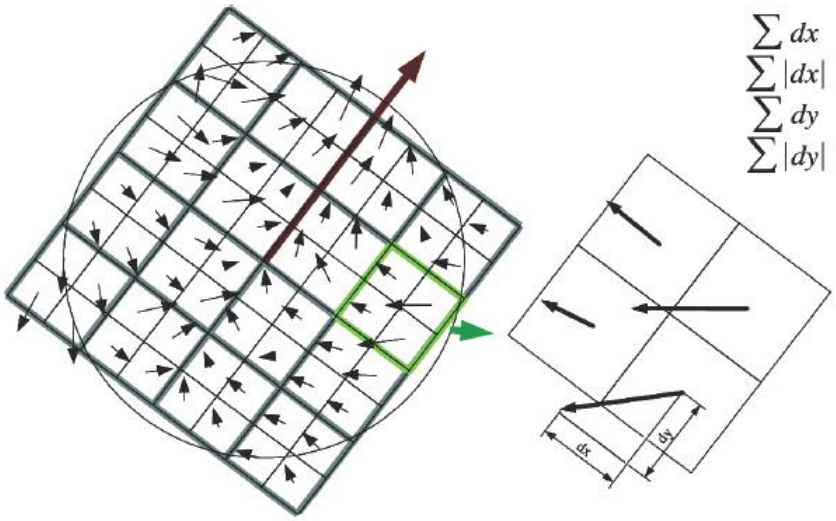

Figura 3.22: Respuestas de Haar en las sub-regiones alrededor del punto de interés

Matching entre puntos clave

Esta sección, al igual que en el caso del descriptor SIFT, representa la correspondencia de los puntos clave identificados entre dos imágenes. La estrategia utilizada para establecer las correspondencias entre los puntos clave de ambas imágenes es la de *"el vecino más próximo"* descrita anteriormente en la sección 3.4.

En el caso del descriptor SURF, el umbral relativo a la estrategia anterior es fijado con un valor de 0,7 [34].

En la Figura 3.23 se representa los resultados del matching de puntos clave entre dos imágenes para el descriptor SURF.

Figura 3.23: Representación del matching para el descriptor SURF

3.6. Ventajas e Inconvenientes de los descriptores utilizados

En esta sección se presenta una tabla a modo de resumen con las ventajas e inconvenientes que, según los autores de los descriptores aquí estudiados, poseen a priori en distintas situaciones o para distintas aplicaciones. Será en las secciones 4.5 y 4.17 donde se presentarán las diferentes ventajas e inconvenientes de los mismos aplicadas al ámbito de este proyecto.

Descriptor	***Ventajas***	***Inconvenientes***
Histograma de Color	Resulta robusto frente a pequeños cambios de escala o pequeños movimientos de los elementos presentes en la imagen. Se muestra invariante respecto de la rotación sobre los ejes. Sencillo y compacto presenta un bajo coste computacional, respecto del tamaño y del tiempo de cálculo.	No incluye información espacial: dos imágenes completamente distintas pueden tener histogramas similares. Las variaciones de iluminación pueden alterar el histograma de forma muy significativa.
Color Layout	Rapidez y repetibilidad. Escasa complejidad del proceso de comparación obteniendo un alto rendimiento respecto del número de comparaciones por unidad de tiempo. Gran compactación de la información que alberga el descriptor.	Efecto bloque de la transformada DCT. Sólo tiene en cuenta la correlación entre los pixels del mismo bloque y no atiende a las relaciones con los pixels vecinos.
Correlograma Color	Incluye información espacial sobre la distribución de los colores de la imagen.	Para una valor de *d(distancia entre pixels)* alto, el coste computacional resulta demasiado elevado.
SIFT	Invarianza respecto de rotaciones, translaciones, escala y cambios de iluminación. Repetibilidad.	Alto coste computacional. Tamaño del descriptor mucho mayor que los anteriores. Discretización de los filtros gaussianos. Dependiente del tamaño de la imagen.
SURF	Mayor robustez y velocidad de cálculo respecto del descriptor SIFT.	Menor tamaño que el descriptor SIFT pero todavía incomparable respecto de los anteriores. Dependiente del tamaño de la imagen.

Tabla 1: Ventajas e inconvenientes de los descriptores

4. Evaluación de Descriptores Aplicados a la Identificación de Imágenes Cuasi-Duplicadas

4.1. Introducción

En este capítulo 4 se detallan las distintas actuaciones y mejoras llevadas a cabo sobre cada uno de los descriptores así como los estudios comparativos y de combinación de descriptores respecto de la tarea principal de este proyecto, la detección de escenas cuasi-duplicadas. Al mismo tiempo se exponen los resultados obtenidos mediante diferentes tablas y representaciones gráficas en cada una de las etapas en las que este proyecto ha sido dividido. La Figura 4.1 representa el diagrama de las diferentes etapas así como de los pasos previos necesarios para la creación del marco comparativo como son el diseño de la base de datos y la elección de los criterios de evaluación.

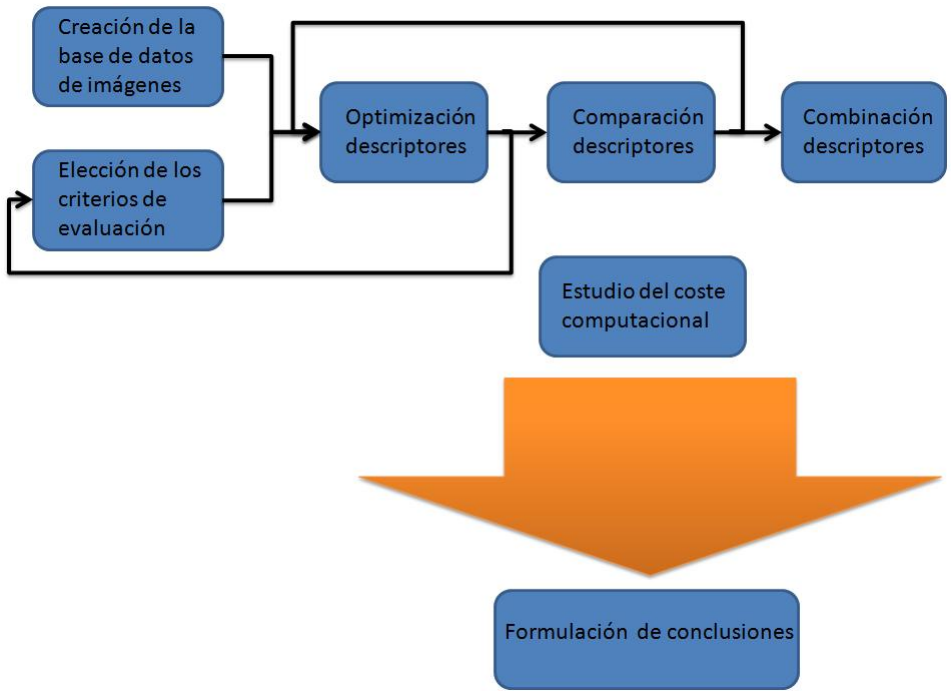

Figura 4.1: Diagrama de las etapas de desarrollo

Antes de describir cada una de estas etapas, se detalla la composición de la base de datos utilizada para el desarrollo del proyecto y se representan diferentes ejemplos de cada una de las categorías. A continuación, se exponen los diferentes criterios de evaluación elegidos para la interpretación de los resultados obtenidos y la formulación de las conclusiones.

En la primera de las etapas, optimización de los descriptores, se evalúa el comportamiento de los descriptores respecto de la tarea de la detección de escenas cuasi-duplicadas teniendo como referencia las características de la base de datos de imágenes. Mediante la variación de los parámetros iniciales y la inserción de modificaciones y/o etapas adicionales, se consigue un incremento en las prestaciones de los descriptores respecto de la formación inicial. Esta etapa de optimización se realiza sobre los descriptores de manera individual.

En la segunda de las etapas, se lleva a cabo un estudio comparativo entre los descriptores ya optimizados con el objetivo de determinar qué descriptores presentan un mejor rendimiento en cada una de las situaciones analizadas, y establecer al mismo tiempo una base comparativa para la siguiente etapa.

En la tercera y última etapa, se realiza un estudio combinativo de los descriptores, persiguiendo mejorar los resultados individuales previos obtenidos en la segunda etapa.

Finalmente se realiza de forma paralela un estudio del coste computacional de los descriptores para resumir los resultados obtenidos mediante conclusiones más precisas, teniendo en cuenta la relación de compromiso entre la precisión de los descriptores para la detección de imágenes relacionadas y el coste computacional que eso requiere.

4.2. Contenido de la Base de Datos

La base de datos de imágenes utilizada ha sido creada específicamente para el propósito de este proyecto a partir de una colección heterogénea de diferentes vídeos de dominio público. Estos vídeos provienen en su gran mayoría de la base de datos para la evaluación de técnicas sobre resúmenes de video del TRECVID[10], en concreto los vídeos sobre series de TV de la cadena inglesa BBC (también conocidos como BBC rushes), así como de los contenidos de Time-Slice® Fiml's videos from Vimeo[11]. En concreto han sido utilizados 38 vídeos procedentes del TRECVID Evaluation con una duración total aproximada de 18 horas y 26 vídeos de 1 hora aproximada de duración procedentes de la segunda colección.

Las imágenes han sido obtenidas a partir de los vídeos mediante una selección manual de acuerdo con los criterios de obtención de los diferentes tipos de escenas analizadas en este proyecto mencionadas en la sección 1.2 y que comparten todas ellas las características semánticas propias de una misma escena como son las imágenes cuasi-duplicadas. El tamaño de las imágenes es de 352 x 288 pixels y ha sido heredado de la resolución de la mayoría de los vídeos, realizando un redimensionamiento en el resto de los casos.

La base de datos de imágenes cuenta con un total de 2000 imágenes distribuidas en 4 grupos o categorías de 500 imágenes cada una, relacionadas con alguno de los cuatro tipos de escenas de estudio en este proyecto: cambios de

[10] http://trecvid.nist.gov/
[11] http://vimeo.com/timeslice

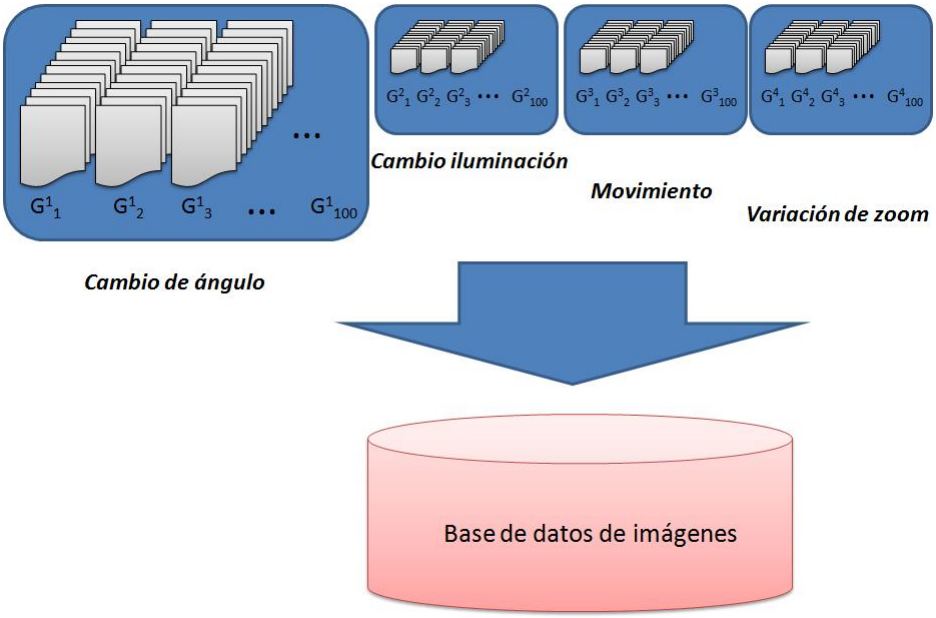

Figura 4.2: Esquema de la base de datos de imágenes

iluminación, de ángulo o punto de vista, movimiento de los elementos de los elementos presentes en la imagen y variaciones de zoom. Cada una de las categorías está a su vez formada por grupos de imágenes que pertenecen a una misma escena pero que están afectadas por alguno de las transformaciones mencionadas anteriormente. El tamaño de estos grupos de imágenes es variable, desde un mínimo de 5 hasta un máximo de 25 imágenes, siendo el tamaño más representativo de 10 imágenes. En la Figura 4.2 se puede ver una representación del esquema completo de la base de datos.

La base de datos al completo utilizada en este proyecto puede ser descargada de la siguiente página web:

http://www-vpu.eps.uam.es/publications/VisualDescriptorComparison/

Las secuencias de imágenes pertenecientes a cambios de ángulo o del punto de vista representan un movimiento de la posición de la cámara con un rango de variación del ángulo desde una vista fronto-paralela a una con una desviación significativa de entre 45°y 50° en ambas direcciones. En la Figura 4.3 se muestra un conjunto de imágenes relacionadas como ejemplo de este tipo de transformaciones.

Los cambios de iluminación son producidos mediante la variación de la posición de la fuente principal de luz en la imagen o mediante la realización de las diferentes tomas bajo diferentes condiciones de iluminación. Un ejemplo de este tipo de transformación fotométrica se puede ver representado en las imágenes de la Figura 4.4.

Figura 4.3: Ejemplo de imágenes afectadas por cambio de ángulo

El movimiento de los elementos presentes en la escena representa uno de los handicap más difíciles en cuanto a la detección de duplicados se refiere. El análisis de una escena en la que la trayectoria de los objetos que la componen no siguen un mismo patrón, conlleva al establecimiento de límites en cuanto al grado de similitud aceptable para poder seguir afirmando que ambas imágenes pertenecen a la misma escena. Es esta categoría la que presenta más diversidad y falta de criterio común a la hora de encontrar una base de referencia para todos los estudios realizados sobre esta problemática con este tipo de escenas.

En el caso analizado en este proyecto, los cambios en la imagen se refieren a pequeños cambios en la posición y movimiento de los objetos tanto en el frente como en el fondo de la imagen. Concretamente, estos cambios suponen a lo sumo una variación, respecto de la composición común de todas las imágenes del grupo relacionado, de un 30 %. En la Figura 4.5 se encuentran representadas diversas imágenes relacionadas con esta problemática.

Por último y respecto de las variaciones de zoom, se identifican dos tipos de variaciones dependiendo a la naturaleza del zoom. La primera de ellas representa los cambios en la distancia entre el objetivo de la cámara y la ubicuidad de la escena. Este tipo de zoom se conoce como zoom físico o acercamiento de la cámara. El segundo de los casos es el conocido como zoom óptico y representa el acercamiento óptico de la escena mediante la variación de las lentes de la cámara, permaneciendo constante la distancia entre cámara y escena. La base de datos utilizada en este proyecto contiene ambos tipos de zoom de forma

Figura 4.4: Ejemplo de imágenes afectadas por cambio de iluminación

aleatoria, y por lo tanto en la presentación de los resultados no se tendrá en cuenta la diferenciación realizada en este apartado, considerándose como un sólo tipo de transformación. En la Figura 4.6 se muestran diversas imágenes relacionadas con este tipo de transformación.

El conjunto de todas las imágenes de la base de datos es dividido en dos grupos: *i) optimization data set,* y *ii) test data set.* El primero de ellos incluye una colección de 10 grupos de imágenes relacionadas por cada una de las categorías de estudio, constituyendo así un conjunto de aproximadamente 400 imágenes (dependiendo del número de imágenes en cada uno de los grupos). La selección de las imágenes ha sido aleatoria para homogeneizar y dotar de una mayor objetividad a la etapa de la optimización de los descriptores mediante la búsqueda de los parámetros óptimos para cada uno de ellos. El segundo de los grupos, *test data set,* está compuesto por la totalidad de las imágenes que forman la base de datos, y será utilizado en ambas etapas de comparación y combinación de descriptores para determinar las diferencias en el rendimiento de éstos tanto de forma individual como combinada respecto de la detección de escenas cuasi-duplicadas.

Figura 4.5: Ejemplo de imágenes con movimiento de los objetos que la componen

Figura 4.6: Ejemplo de imágenes afectadas por variación de zoom

4.3. Sistema de Evaluación

Los resultados del estudio comparativo sobre los descriptores visuales presentados en este proyecto se engloban dentro de la problemática sobre decisiones binarias que es compartida por muchas otras disciplinas como son los mencionados sistemas CBIR o la recuperación de imágenes (image retrieval, IR).

Existen diferentes criterios basados en el número de aciertos y de fallos respecto de las correspondencias establecidas entre dos imágenes y relativos a la problemática mencionada. Provost *et al.* [39] recomienda el uso de curvas Reciever Operator Characteristic (curvas ROC) para evaluar los diferentes sistemas desarrollados en torno a la comparación de imágenes. Sin embargo, las curvas ROC pueden resultar al mismo tiempo erróneamente optimistas cuando la distribución de las imágenes de la base de datos no está lo suficientemente compensada. Drummond *et al.* [40, 41] recomiendan el uso de curvas de costes para tratar estos problemas.

Para la mayor parte de los investigadores cuyos resultados comparten la problemática aquí expuesta, resulta más común el uso de otro sistema de evaluación que también se basa en las diferentes tasas de aciertos y fallos respecto de las correspondencias establecidas. Las curvas *precisión-Recall* (PR) también representan una alternativa a las curvas ROC [42, 43, 44][45], estableciéndose además una relación clara entre ambos criterios como puede apreciarse en el trabajo de Davis *et al.* [46].

Mediante las ecuaciones 4.1 y 4.2 se detallan las componentes de las gráficas PR:

$$precision = \frac{\#\, correct\ matches}{\#\, correct\ matches + \#\, false\ matches} \tag{4.1}$$

$$recall = \frac{\#\, correct\ matches}{\#\, correspondences} \tag{4.2}$$

Tanto el valor de *precision* como de *recall* están relacionados implícitamente con un ranking o posición ordenada de las imágenes objetivamente relacionadas o pertenecientes a la misma escena con respecto al total de las imágenes que componen el grupo de comparación o análisis en cada momento.

Dada la composición y singularidad de la base de datos utilizada donde las imágenes se encuentran distribuidas en grupos interrelacionados, se considerarán como *correct matches*, aquellas imágenes comparadas que, perteneciendo a la misma escena o grupo interrelacionado que la imagen original dentro de la categoría analizada, son identificadas de forma acertada como imágenes relacionadas.

De la misma manera, el término *false matches* corresponde a las imágenes que tras el proceso de comparación han sido identificadas como relacionadas con la imagen original de forma incorrecta, pues no pertenecen a la misma escena.

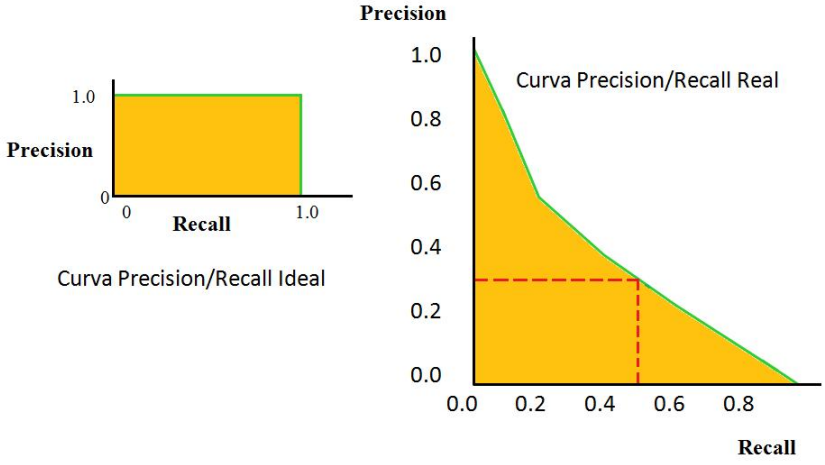

Figura 4.7: Ejemplo de curvas PR

Por ultimo el número de correspondencias representa el número de imágenes que están relacionadas con la imagen original, es decir, que pertenecen a la misma escena.

De esta manera a medida que el cardinal del grupo de imágenes comparadas crece y se acerca al número de imágenes totales, el número de imágenes identificadas como matching correctos tiende al número de imágenes totales pertenecientes a la misma escena que la imagen original.

Antes de presentar los resultados en las siguientes secciones, se explican aquí las posibles curvas y formas que pueden adoptar las gráficas de PR. La Figura 4.7 muestra un ejemplo teórico de una de estas gráficas con el objetivo de resultar más explícitos en cuanto a los detalles de las curvas.

El comportamiento ideal de un descriptor presenta un valor de *precision* igual a 1 para cualquiera que sea el valor de *recall*, lo que se traduce en una curva horizontal y constante alcanzando la esquina superior derecha, donde el descriptor es capaz de realizar el matching entre todas las imágenes relacionadas con la misma escena y la imagen original. En la práctica, la precisión del descriptor decrece debido a un incremento en la distancia entre los descriptores similares como consecuencia de las transformaciones sobre las imágenes. Un caída abrupta de la curva PR es una muestra de que el descriptor sufre de la degradación de la imagen (cambios de iluminación, zoom, etc.). Finalmente, si las curvas pertenecientes a diferentes descriptores están muy separadas entre sí o tienen distintas pendientes, significa que el carácter distintivo y la robustez de ambos descriptores es diferente respecto de la transformación o tipo de escena analizada.

Dado que las gráficas PR no presentan de manera explícita toda la información deseada, se utilizan otras medidas basadas en el *precision* y el *recall* entre las cuales se utiliza en este proyecto el dato del área bajo la curva de PR (AUC-PR), como una métrica simple para determinar el comportamiento de los diferentes descriptores y obtener un criterio simple de comparación entre los mismos [46].

Para llevar a cabo la tarea de evaluación y el cálculo tanto de las gráficas PR como el valor de AUC-PR se ha utilizado un programa en lenguaje Java de uso público llamado *AUCCalculator* y que se encuentra disponible para su descarga en la siguiente pagina web: http://mark.goadrich.com/programs/AUC/

Este programa ha sido desarrollado por los mismos autores del trabajo anteriormente referenciado como Davis *et al.* [46].

4.4. Optimización de Descriptores: Análisis intra-descriptor

En esta sección se presentan las diferentes modificaciones y elección de parámetros sobre cada uno de los descriptores, así como la utilización de varias métricas para el cálculo de la distancia[12] o *score* entre las imágenes comparadas. De esta manera se destaca la mejor combinación de parámetros para cada descriptor en base a los resultados obtenidos y que será utilizada en la siguiente etapa de comparación de descriptores.

A continuación se presentan las diferentes optimizaciones para 4 de los 5 descriptores del proyecto, exceptuando en esta etapa la optimización del descriptor Color Layout, Esta decisión obedece al hecho de que el descriptor pertenece al estándar MPEG-7 y que por lo tanto se ha decidido contar con él de la manera en que está originalmente especificado

Para llevar a cabo las diferentes optimizaciones de los descriptores se ha utilizado el conjunto de imágenes denominado *optimization data set*. La manera de realizar las comparaciones entre las imágenes se resume de la siguiente manera:

Cada uno de los descriptores es evaluado mediante la comparación de diferentes imágenes de referencia, denominadas *imágenes query*, con un conjunto más amplio de imágenes. La elección tanto de las *imágenes query* como de los grupos de comparación respectivos de cada query se realizan de forma aleatoria, obteniendo 10 *imágenes query* por cada una de las categorías analizadas, y un grupo diferente de 150 imágenes de entre todas las imágenes de la base de datos para cada una de las *querys*. En cada uno de los grupos de comparación de imágenes se encuentran todas las imágenes relacionadas con la *query* al que son asignados, y el resto son elegidas de forma aleatoria.

En resumen, el *optimization data set* cuenta con 40 *imágenes query* con sus 40 grupos respectivos de 150 imágenes cada uno con las que efectuar las comparaciones.

[12]Los términos distancia y score se utilizan a partir de este momento de forma indistinta para referirse al mismo concepto de cuantificación de la similitud entre dos imágenes

4.4.1. Histogramas de Color RGB y HSV

Como se ha mencionado en el sección 3.1 se han utilizado dos espacios de color diferentes para el descriptor histograma de color, resultando de esta manera, la implementación de dos descriptores que, a efectos de presentación de resultados y elaboración de conclusiones se considerarán como dos descriptores independientes.

Las implementaciones de ambos descriptores, histogramas de color RGB y HSV, se han realizado personalmente mediante la adaptación de la teoría expuesta en la sección 3.1. En ellas se han llevado a cabo algunas decisiones de diseño en congruencia con las recomendaciones y resultados mostrados en [47], que si bien podrían haber formado parte de parámetros adicionales, se ha preferido dejarlo como trabajo futuro. Estas decisiones se refieren al número de intervalos o *bins* en los que se divide el histograma y que se han fijado en 32 para las tres componentes de color en el caso del histograma RGB y en 16, 4, 4 para las componentes hue, saturación y value en el caso del histograma HSV. Por lo tanto se obtienen los histogramas de cada una de las componentes y que a su vez se concatenan de manera que el descriptor identifica las tres componentes de color en un mismo vector concatenado.

La optimización propiamente dicha de los histogramas se ha llevado a cabo mediante dos modalidades de creación y comparación de los descriptores así como de diferentes métricas implementadas para el cálculo de la distancia entre las imágenes comparadas.

Las modalidades de histograma son:

- histograma de la imagen completa
- histograma de la imagen por regiones: las regiones resultan de dividir la imagen en 4 cuadrículas, obteniendo así un histograma por cada región. De la misma manera, se considera la partición en un número de regiones diferente o la forma alternativa de las regiones como posible trabajo futuro.

La comparación entre los descriptores de las imágenes se realiza bien mediante el histograma de la imagen completa o bien mediante la comparación de los histogramas de cada cuadrícula respectivos de cada imagen.

Respecto de las métricas implementadas, se han elegido 5 distancias diferentes conocidas y utilizadas ampliamente en el cálculo de distancias entre vectores como son: Bhattacharyya, Cityblock, Euclidea, Intersección y Chi-cuadrado. Las fórmulas que las resumen son las siguientes:

Definition 1. Dados dos vectores $x = (x_1, x_2, ..., x_n)$ e $y = (y_1, y_2, ..., y_n)$ se definen las siguientes distancia:

$$Bhattacharyya = -\ln\left(\sum_{i=1}^{n}(\sqrt{x_i} * \sqrt{y_i})\right) \tag{4.3}$$

$$Cityblock = \sum_{i=1}^{n} \mid x_i - y_i \mid \tag{4.4}$$

$$Euclidea = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4.5}$$

$$Intersección = 1 - \frac{\sum_{i=1}^{n} \text{mín}\,(x_i, y_i)}{\text{mín}\,(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i)} \tag{4.6}$$

$$Chi - cuadrado = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{2 * (x_i + y_i)} \tag{4.7}$$

Las Tabla 2 muestra los resultados obtenidos para cada una de las modalidades y distancias implementadas respecto de las diferentes categorías de escenas analizadas. Además de las 4 categorías ya mencionadas la evaluación se concreta también para un tipo de *escenario global* que resulta ser un promedio de los rendimientos mostrados en las otras 4 categorías y que intenta modelar la existencia de todas ellas de forma conjunta.

Observando detenidamente los datos mostrados en las tablas se puede apreciar como para ambos descriptores, histograma RGB y HSV, la métrica que mejor rendimiento obtiene es la distancia Cityblock, también conocida como distancia L_1, en la modalidad de comparación por regiones. Si bien no alcanza los mejores resultados en todos y cada una de los tipos de escena analizados, como es el caso del cambio de ángulo, si lo hace en la mayoría de las categorías y por lo tanto en el tipo de escena global. Se observa en las tablas de resultados de ambos descriptores como por un lado la distancia Euclídea reporta el peor de los rendimientos obtenidos y al mismo tiempo se aprecia que las métricas calculadas a partir de regiones alcanzan por lo general mejores resultados que las que operan teniendo en cuenta la totalidad de la imagen. También es destacable que el rendimiento alcanzado por parte de la métrica *Cityblock* no supone una gran ventaja respecto de otras métricas como puede ser el caso de la métrica *Intersección de histogramas* por regiones.

Resulta necesario dejar constancia de que en esta sección de optimización de descriptores se han de elegido los parámetros que obtienen un mayor rendimiento medio, para la satisfacción de la mayor parte de las situaciones reales.

Teniendo en cuenta todas estas apreciaciones, se concluye que la elección de la métrica óptima en el caso de ambos descriptores, y por tanto la que se utiliza en las etapas posteriores, es la métrica o distancia $Cityblock_2$ o por regiones.

Tipo de escena - Ranking		1º	2º	3º	4º	5º	6º	7º	8º	9º	10º
Cambio de ángulo	Tipo distancia	B_1	Cb_1	I_1	Cs_1	I_2	Cb_2	Cs_2	B_2	E_1	E_2
	Valor	0.9663	0.9492	0.9487	0.9464	0.9438	0.9403	0.9378	0.9343	0.9056	0.8912
Cambio Iluminación	Tipo distancia	Cb_2	I_2	Cs_2	B_2	B_1	Cb_1	I_1	Cs_1	E_1	E_2
	Valor	0.7718	0.7689	0.7637	0.7563	0.7359	0.7348	0.7317	0.7317	0.6709	0.6007
Escena de Movimiento	Tipo distancia	Cb_2	Cs_2	I_2	B_2	Cs_1	B_1	Cb_1	I_1	E_2	E_1
	Valor	0.8444	0.8433	0.8413	0.8173	0.8138	0.8115	0.8093	0.8088	0.7751	0.7592
Variación de Zoom	Tipo distancia	Cb_2	I_1	Cb_1	I_2	Cs_1	Cs_2	B_1	B_2	E_1	E_2
	Valor	0.8357	0.8357	0.8343	0.8342	0.8192	0.81	0.8091	0.7906	0.7793	0.7422
Escena Global	Tipo distancia	Cb_2	I_2	Cs_2	Cb_1	I_1	B_1	Cs_1	B_2	E_1	E_2
	Valor	0.8477	0.8411	0.8387	0.8323	0.8312	0.8307	0.8278	0.8246	0.7787	0.7523

a)

Tipo de escena - Ranking		1º	2º	3º	4º	5º	6º	7º	8º	9º	10º
Cambio de ángulo	Tipo distancia	B_1	Cb_1	I_1	Cs_1	Cb_2	I_2	E_1	Cs_2	E_2	B_2
	Valor	0.9801	0.965	0.9645	0.9638	0.9633	0.9631	0.9605	0.9544	0.9541	0.9337
Cambio Iluminación	Tipo distancia	Cb_2	I_2	Cs_2	B_1	I_1	Cb_1	E_2	B_2	Cs_1	E_1
	Valor	0.8343	0.8325	0.8316	0.8143	0.814	0.8133	0.8115	0.8115	0.8107	0.7848
Escena de Movimiento	Tipo distancia	I_1	Cb_2	B_1	I_1	Cs_2	I_2	Cs_1	B_1	E_1	B_2
	Valor	0.8581	0.8549	0.8528	0.8504	0.8482	0.8457	0.8429	0.842	0.8395	0.8374
Variación de Zoom	Tipo distancia	Cs_2	Cb_2	I_2	Cs_1	Cb_1	I_1	B_2	I_2	B_1	E_1
	Valor	0.9055	0.9028	0.9015	0.8919	0.8849	0.8835	0.8792	0.8731	0.8684	0.8586
Escena Global	Tipo distancia	Cb_2	I_2	Cs_2	B_1	I_1	Cs_1	Cb_1	E_2	B_2	E_1
	Valor	0.889	0.882	0.8806	0.8789	0.8781	0.8773	0.8763	0.8711	0.8654	0.8608

b)

Tabla 2: Optimización histograma a) RGB y b) HSV

El valor del área bajo la curva de PR para las diferentes métricas Bhattacharyya (B), Chi-square (Cs), Cityblock (Cb), Euclidea (E) e Intersección (I), con los subíndices 1 y 2 para diferenciar la aplicación de la métrica sobre la imagen completa o por regiones respectivamente.

4.4.2. Color Layout

A pesar de que el descriptor Color Layout está exento de la etapa de optimización, se presenta en este apartado la elección del número de coeficientes de la DCT para la creación del descriptor que difiere del valor establecido por defecto.

Según el estándar, tanto para la creación del descriptor como para las posteriores comparaciones de imágenes, solo se tienen en cuenta 6 coeficientes para la componente Y y 3 para las componentes Cr y Cb por defecto, aunque es posible la elección de 1, 3, 6, 10, 15, 21, 28 y 64 coeficientes. Si bien esta recomendación se basa en la apenas ausencia de diferencias en los resultados respecto de las elecciones de un mayor número de coeficientes, para el caso aquí presentado se tienen en cuenta los 64 coeficientes.

Por otro lado, respecto del valor de los pesos correspondientes a la ecuación 3.7 sobre la métrica de comparación para este descriptor, en el estándar se establece que la ponderación de los coeficientes en las tres componentes de color sean decrecientes respecto del orden de escaneado de los coeficientes, dotando de esta manera con un mayor pero a los coeficientes de bajas frecuencias. Además la ponderación también tiene en cuenta la importancia diferente con respecto a la componente de color. En el caso aquí presentado se ha optado por la elección de un índice de pesos acorde a la posición y que tiene en cuenta la diferencia entre componentes como se aprecia en las siguientes fórmulas:

$$w_i^Y = \frac{0{,}2}{(j+k)^2}, \quad w_i^{Cb} = \frac{0{,}4}{(j+k)^2}, \quad w_i^{Cr} = \frac{0{,}4}{(j+k)^2}$$

donde i representa el índice del coeficiente según el orden de escaneado y (j,k) representan los índices dentro de la matriz de coeficientes de la DCT.

4.4.3. Correlograma

Optimización en la etapa de creación del descriptor:

Al igual que en el caso de los descriptores anteriores, la implementación del correlograma de color utilizada en este proyecto surge de la adaptación de los conceptos teóricos sobre el mismo.

Como se ha mencionado en la sección 3.3, el calculo del correlograma de color es interpretado como una relación entre los diferentes colores y la disposición espacial que existe entre ellos en la imagen. Basándose en los resultados y conclusiones expuestas en [26], se lleva a cabo el cálculo del correlograma teniendo en cuenta tan sólo la disposición espacial o relación entre colores similares de la imagen, lo que se conoce como auto-correlograma.

En referencia a la definición del correlograma expresada en la ecuación 3.8, el auto-correlograma se define mediante:

$$\alpha_c^k(I) = \gamma_{c,c}^k(I) \tag{4.8}$$

Esta simplificación del concepto más amplio del correlograma reduce la capacidad discriminativa del mismo en una proporción bastante inferior respecto de la reducción de tamaño mucho más considerable y cuya relación de compromiso resulta ser beneficiosa. Han sido éstas las razones por las que se ha optado en este proyecto por la utilización del auto-correlograma en detrimento del correlograma de color en su sentido más completo.

Durante la implementación también se han tomado diferentes decisiones que afectan al rendimiento del descriptor. Dado que el correlograma representa la información de la relación espacial entre colores similares mediante la creación de histogramas, se ha establecido una cuantificación de 16 bins para cada una de las componentes de color RGB de la imagen, de manera que los colores de la imagen se ven cuantificados mediante la representación de 16 valores de tonalidad para cada una de las componentes. Como resultado se obtienen 3 correlogramas como representación de cada imagen. La elección del sistema de color RGB viene determinada por la descripción teórica del propio descriptor, indicando otras posibles representaciones de color como trabajo futuro.

La parametrización del correlograma queda representada mediante la variación de las distancias para el cálculo del correlograma y mediante la diversificación de las proporciones de la imagen respecto del las dimensiones originales de la misma.

La distancia sobre la cual se realiza el cálculo de las relaciones de similitud entre los pixels de la imagen queda representada mediante la ecuación 3.8 por el valor de k. La parametrización de este valor se lleva a cabo mediante un vector con 5 componentes de distancias: $v = (1, 3, 5, 7, 9)$. Para cada una de las componentes, se calcula un correlograma restringiendo la búsqueda de colores similares a un radio del valor de cada componente. De esta manera por ejemplo, se obtiene un correlograma que representa la disposición espacial de los colores similares a una distancia máxima de 5 pixels. El vector de distancias es suficientemente grande como para establecer un margen amplio de variabilidad y con un paso suficientemente pequeño para representar los pequeñas transiciones de color presentes en las imágenes.

Con respecto al valor del tamaño de la imagen en la práctica respecto de la original, se han establecido 4 posibles valores representados esta vez por un vector de redimensionamiento $r = (1, 200, 100, 50)$. Es necesario recordar que el tamaño de las imágenes de la base de datos es de 352 x 288 pixels, y por lo tanto, los valores de redimensionamiento afectan al tamaño de la imagen que se utiliza para el cálculo del correlograma según:

- 1 $\rightarrow\longrightarrow$ Tamaño original de la imagen.
- 200 $\longrightarrow$ Imagen cuadrada de 200 x 200 pixels
- 100, 50 $\longrightarrow$ Ambos casos similar al caso de 200 pero con los valores de 100 y 50 respectivamente.

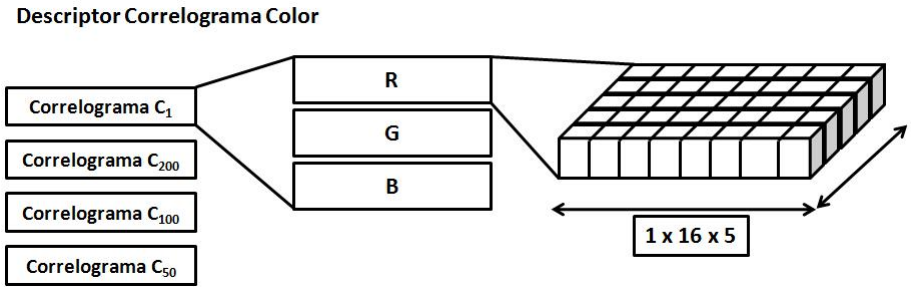

Figura 4.8: Esquema del descriptor correlograma de color

Como resultado se obtiene un descriptor que está representado por una estructura con 4 campos, uno para cada componente del vector de redimensionamiento r. En cada uno de los campos se aloja el correlograma de la imagen del tamaño correspondiente que a su vez se encuentra representado mediante 3 variables, una por cada componente de color, y dentro de cada una de las cuales se encuentra nuevamente una matriz de 3 dimensiones. Dos de las dimensiones de la matriz almacenen el valor los histogramas de las distribuciones espaciales de los colores, mientras que la 3 dimensión representa la componente del vector de distancias v utilizada para el cálculo de los correlogramas. La Figura 4.8 muestra un esquema de la composición del descriptor para la ayuda de su comprensión.

Optimización en la etapa de comparación del descriptor:

En el proceso de comparación de las imágenes se diferencian los resultados atendiendo a los diferentes valores de redimensionamiento, es decir, se clasifican los resultados realizando los cálculos para cada componente del vector r por separado. De esta manera se modela el comportamiento del descriptor mediante la parametrización mencionada anteriormente. Una vez seleccionado el valor de la componente r en cuestión se realiza la comparación entre dos imágenes mediante la diferenciación individual de los correlogramas de cada una de las componentes de color RGB y de cada una de las distancias. Los tres valores procedentes de la diferenciación son ponderados de manera similar para componer un único valor de *score* que identifique la similitud entre las dos imágenes.

La Tabla 3 muestra los resultados obtenidos en esta etapa de optimización diferenciando los casos para los diferentes valores de redimensionamiento de la imagen.

A tenor de los resultados mostrados en la tabla, se puede observar como el redimensionamiento de la imagen afecta negativamente al rendimiento del descriptor para todas las categorías o tipos de escena analizados excepto en el caso de los cambios de iluminación. Lo primero es debido a que el redimensionamiento de la imagen comprime la información mediante el agrupamiento de los pixels en un único valor, produciendo una pérdida de información irreversible. El mo-

Tipo de escena - Ranking		1º	2º	3º	4º
Cambio de ángulo	Dimensiones	C_1	C_{200}	C_{100}	C_{50}
	Valor	0.9597	0.9336	0.9261	0.9144
Cambio Iluminación	Dimensiones	C_{50}	C_1	C_{200}	C_{100}
	Valor	0.6939	0.6851	0.6842	0.6808
Escena de Movimiento	Dimensiones	C_1	C_{200}	C_{100}	C_{50}
	Valor	0.7989	0.7976	0.7864	0.7727
Variación de Zoom	Dimensiones	C_{200}	C_1	C_{100}	C_{50}
	Valor	0.7635	0.748	0.7398	0.7333
Escena Global	Dimensiones	C_1	C_{200}	C_{100}	C_{50}
	Valor	0.798	0.7917	0.7833	0.7786

Tabla 3: Optimización Correlograma

tivo por el cual se había parametrizado el descriptor mediante este valor tenía que ver con la búsqueda de un rendimiento mayor en cuanto al coste computacional, que si bien se consigue, no supone tras la evaluación de los resultados, una mejora respecto de la pérdida de rendimiento. La selección de un tamaño de imagen menor podría ser interesante en el caso de trabajar con imágenes de mayor tamaño ya que la relación entre la pérdida de aproximadamente un 2 % de rendimiento podría estar más compensada con una mayor reducción del coste computacional.

Si bien es cierto que no existen diferencias relevantes en ninguna de las categorías analizadas, resulta curioso el caso de la inversión en el rendimiento del descriptor para el caso de los cambios de iluminación. Este resultado parece obedecer al hecho de que estos cambios afectan directamente y de forma muy pronunciada al valor de los pixels. De esta manera existen más diferencias entre la imagen original y la transformada y por consiguiente el agrupamiento de los pixels mediante el redimensionamiento amortigua este defecto, obteniendo así mayores rendimientos. De nuevo se hace incapié en que a tenor de las escasas diferencias obtenidas en los resultados respecto de los diferentes valores de redimensionamiento, no es posible establecer una conclusión clara al respecto. Sin embargo, ante la necesidad de elegir una de las opciones y en base a los datos mostrados en la tabla, es la versión C_1 del correlograma la elegida para los estudios de las etapas posteriores.

4.4.4. SIFT

El algoritmo encargado de modelizar el comportamiento del descriptor SIFT utilizado en este proyecto está formado por diferentes funciones implementadas en el lenguaje de programación Matlab®. La autoría del código pertenece al Andrea Vedaldi y forma parte de Vision Lab - Department of Computer Science of University of California, UCLA[13].

Haciendo referencia a la sección 3.4en la que se explicaba el desarrollo del descriptor SIFT, aparecen ciertos parámetros que han sido establecidos por defecto y que siguen las recomendaciones del paper de referencia para el descriptor SIFT [8]. Estos valores, se recuerda, son los siguientes:

- $\sigma = 1{,}6 \longrightarrow$ factor de escala gaussiano.
- $s = 3 \longrightarrow$ relativo al número de imágenes en cada octava.
- $D = 0{,}03 \longrightarrow$ umbral relativo a la eliminación de puntos clave con bajo contraste en el proceso de localización de puntos clave estables.
- $r = 10 \longrightarrow$ relativo a la eliminación de puntos clave en los bordes dentro del proceso de localización de puntos clave estables.
- $\frac{d_1}{d_2} < 0{,}76 \longrightarrow$ ratio relativo a la etapa de matching entre puntos clave.

Estos parámetros son recogidos en el código de la implementación y a su vez se especifican otros que tienen que ver con la versión particular del mismo:

- $FirstOctave = -1 \longrightarrow$ El valor -1 tiene el efecto de duplicar el tamaño de la imagen antes del cálculo del espacio-escala.
- $NumberOfOctaves = \perp(\log(min(M, N))) - FirstOctave - 3 \longrightarrow$ número de octavas calculadas en el espacio-escala gaussiano, donde M y N representan las dimensiones de la imagen.

La optimización propia del descriptor SIFT presentada en esta sección está asociada solamente a la etapa de matching entre puntos clave. En concreto, el proceso de optimización se realiza mediante la implementación de 3 métricas diferentes, a su vez con distintas parametrizaciones, que realizan el cálculo de las correspondencias establecidas entre ambas imágenes; la inserción de una etapa adicional de filtrado de correspondencias entre puntos clave previamente calculadas; y finalmente, mediante la ponderación parametrizada de ambas componentes, espacial y calidad de las correspondencias, que resulta en un valor final de la distancia que representa el grado de similitud existente entre ambas imágenes comparadas, y que de forma posterior se utilizará para representar los resultados de forma ordenada en función del parecido entre las imágenes.

Resulta necesario dejar constancia de que tanto la segunda como la tercera de las mejoras mencionadas no forman parte de la descripción del método original

[13] El código puede ser descargado de la pagina web: http://www.vlfeat.org/~vedaldi/code/sift.html

del descriptor SIFT y que en este proyecto la segunda de las mejoras recibe el nombre de *eliminación de falsas correspondencias*.

Métricas utilizadas

La etapa de matching entre puntos clave del descriptor SIFT es optimizada en una primera aproximación mediante el uso de 3 métricas diferentes que establecen las correspondencias entre ambas imágenes comparadas:

Definition 2. C_1: Denominada *distancia umbralizada*, esta distancia establece la correspondencia entre dos puntos clave cuando la distancia euclídea entre sus vectores de características es menor que un cierto umbral, $threshold_1$ (parámetro), y además esta distancia es mínima respecto de la totalidad de los puntos clave de la imagen comparada. Como restricción añadida se establece también que las correspondencias entre puntos clave sean únicas, es decir, que no existan puntos que tienen más de una correspondencia establecida. En caso de tener más de una correspondencia, sólo quedará como definitiva aquella cuya distancia euclídea sea la mínima entre las correspondencias de ese punto. Esta última parte establece una biyección entre los puntos clave de ambas imágenes.

Definition 3. C_2: Denominada *distancia basada en el vecino más próximo*, establece el matching entre aquellos puntos clave de ambas imágenes que cumplen lo siguiente: De entre las distancias euclídeas entre el punto clave A de la imagen fuente y todos los puntos clave de la imagen comparada, se eligen las dos distancias mínimas d_A y d_B que corresponden a los puntos B y C. Si se cumple que el ratio entre ambas distancias es menor que un cierto umbral $threshold_2$ (parámetro), $\frac{d_A}{d_B} < threshold_2$, junto con la restricción de biyectividad expuesta en el punto anterior, entonces se realiza el matching entre los puntos A y B de las imágenes fuente y comparada respectivamente.

Definition 4. C_3: Denominada *distancia mínima simple*, registra las correspondencias entre los puntos clave de las imágenes cuya distancia euclídea entre ellos sea mínima. También se aplica la restricción de biyectividad anterior.

Una vez expuestas las 3 estrategias de matching entre puntos clave, se presentan a continuación los diferentes valores utilizados en esta etapa de optimización para los dos parámetros identificados anteriormente:
$threshold_1 = (0{,}3,\ 0{,}5,\ 0{,}7,\ 0{,}8)$
$threshold_2 = (0{,}5,\ 0{,}7)$

Como resultado de esta primera sub-etapa de optimización se obtienen las correspondencias o matching entre los puntos clave, identificando su posición dentro de la imagen a la que pertenecen y el índice que ocupan en el emparejamiento respecto del índice natural que tenían los puntos antes del emparejamiento.

Eliminación de Falsas Correspondencias

Para medir la discriminación de las características es importante saber cuánto de fiable puede llegar a ser el proceso de correspondencia entre los puntos clave de ambas imágenes. El procedimiento de correspondencia entre estos puntos puede generar algunos emparejamientos erróneos. Además es posible que puntos pertenecientes a una de las imágenes no tengan correspondencia con ningún punto de la imagen de comparación debido a oclusiones o a que no sean detectados.

Existen diferentes aproximaciones para resolver este problema [48, 49]. Una de las más intuitivas consiste en eliminar las correspondencias espúreas utilizando restricciones geométricas limitando las posibles variaciones geométricas a pequeñas rotaciones y desplazamientos dentro de la imagen [48]. De esta manera, si se colocan dos imágenes objetivamente relacionadas de forma consecutiva y se visualizan las correspondencias establecidas entre los puntos de ambas imágenes como puede apreciarse en la Figura 4.9 (arriba), las correspondencias verdaderas deben aparecer como lineas paralelas con una longitud similar.

De acuerdo con esta observación, se calcula una orientación θ_P y una longitud l_P predominantes respecto de todas las correspondencias establecidas, y a continuación se realiza un filtrado de las mismas seleccionando aquellas cuya orientación θ y longitud l estén dentro de unos márgenes preestablecidos ϵ_θ y ϵ_l de manera que:

$$\mid \theta - \theta_P \mid < \epsilon_\theta \tag{4.9}$$

$$\mid l - l_P \mid < \epsilon_l \tag{4.10}$$

Es necesario reseñar que este proceso de filtrado de correspondencias entre puntos clave no está incluido dentro de la descripción original del descriptor SIFT [8], si bien puede encontrarse en otras implementaciones orientadas a aplicaciones [48]. Este método obtiene buenos resultados debido a que las correspondencias correctas entre puntos clave deben de tener a su vecino mas parecido más cerca que a el más parecido de las correspondencias incorrectas para llevar a cabo un proceso de correspondencia entre puntos clave de garantía.

Con respecto al cálculo de la orientación y longitud predominantes, θ_P y l_P respectivamente, se ha realizado un estudio estadístico para determinar sus valores. Concretamente ambas variables son modeladas mediante una función de distribución de probabilidad normal o también llamada gaussiana. La distribución de una variable normal está completamente determinada por dos parámetros, su media μ y su desviación estándar σ y cuya función queda determinada por la ecuación 4.11:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{-1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\}; \quad -\infty < x < \infty \tag{4.11}$$

donde,

$$\theta_P = \mu = \frac{1}{n} \sum_{i=1}^{n} \theta_i \tag{4.12}$$

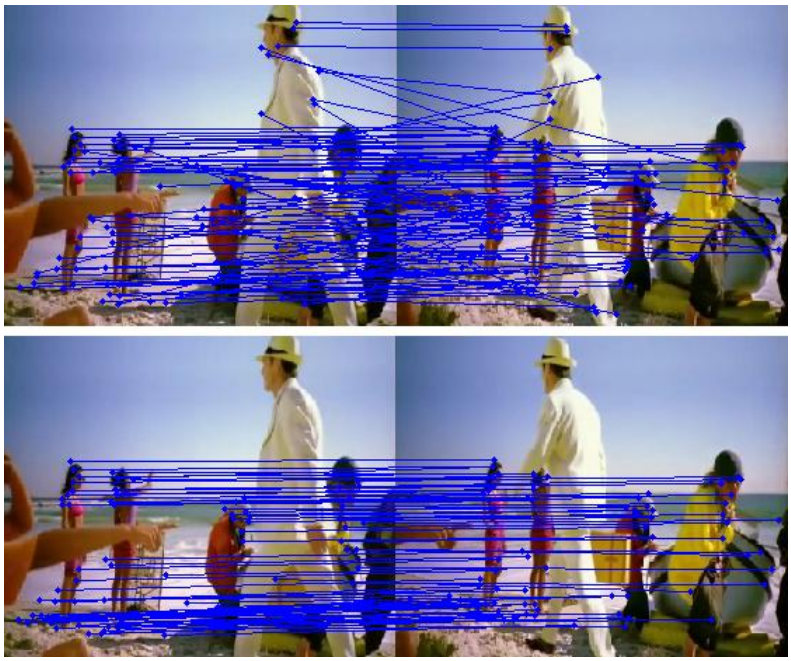

Figura 4.9: Eliminación de correspondencias espúreas

$$\sigma_\theta = \sqrt{\frac{\sum_{i=1}^{n} (\theta_i - \mu_\theta)^2}{n-1}} \tag{4.13}$$

para el caso de la orientación y $l_P = \mu$ y σ_l para la longitud.

De esta manera se obtienen los valores de la orientación y la longitud predominantes así como las desviaciones típicas muestrales respecto de todas las correspondencias establecidas entre los puntos clave.

El filtrado de las correspondencias establecidas se produce mediante la aplicación de las ecuaciones 4.9 y 4.10 estableciendo el valor de los márgenes de paso del filtro, $\epsilon_\theta = 4\sigma_\theta$ y $\epsilon_l = 4\sigma_l$ de forma experimental.

El resultado de este proceso puede verse en la Figura 4.9 (abajo) apreciándose la ausencia de lineas que se cruzan entre ambas imágenes y una longitud entre los puntos matcheados uniforme.

Calculo del *Score*

La etapa de optimización se completa mediante el cálculo de la distancia o *score* que representa una medida de similitud entre las dos imágenes comparadas. Para ello se hace uso de los resultados obtenidos de las correspondencias establecidas por las dos sub-etapas anteriores. Mediante la ecuación 4.14 se calcula la relación entre el numero de puntos clave detectados en las imágenes y el

número de emparejamientos establecidos. Como resultado se obtiene una distancia $D_{matching} \in [0,1]$ cuya interpretación resulta en un análisis del porcentaje de puntos clave que han resultado emparejados. Esta distancia tan sólo tiene en cuenta el *número* de correspondencias entre ambas imágenes.

Por otro lado, mediante la ecuación 4.15 se calcula la relación espacial extraída de la posición de los puntos clave emparejados. Este tipo de distancia guarda relación con la calidad o geometría de las correspondencias establecidas.

$$D_{matching} = \frac{\left(\frac{NumMatches}{NumKeypoints_{Im1}}\right) + \left(\frac{NumMatches}{NumKeypoints_{Im2}}\right)}{2} \tag{4.14}$$

$$D_{spatial} = \frac{1}{NumMatches} \sum_{i=1}^{NumMatches} \sqrt{\left(\left(\frac{y_i^{Im1} - y_i^{Im2}}{M}\right)^2 + \left(\frac{x_i^{Im1} - x_i^{Im2}}{N}\right)^2\right)} \tag{4.15}$$

donde M y N representan las dimensiones de ambas imágenes y x_i^{Im1}, y_i^{Im1}, x_i^{Im2} e y_i^{Im2} representan la posición x e y de los puntos clave matcheados en las imágenes a las que pertenecen.

Estas dos distancias son ponderadas mediante diferentes pesos identificados como w_1 y $w_2 = 1 - w_1$, $(w_1, w_2) \in [0,1]$ (parámetros) en la ecuación 4.16. El valor de los parámetros varía no solo como función de la propia búsqueda de la optimización del rendimiento del descriptor, sino también dependiendo de la estrategia de matching C_1, C_2, C_3 elegida. De esta forma se exponen a continuación el valor de los pesos en función de la estrategia seleccionada.

$$Score = (w_1\,(1 - D_{matching})) + (w_2 D_{spatial}) \tag{4.16}$$

En el caso de C_1: w_1 = (0,33, 0,43, 0,5, 0,6, 0,66, 0,73);
En el caso de C_2: w_1 = (0,65, 0,6, 0,7, 0,75, 0,8);
En el caso de C_3: w_1 = (0, 0,1, 0,3, 0,5);

Finalmente se exponen en la Tabla 4 los resultados obtenidos en la etapa completa de optimización para el descriptor SIFT. Para poder interpretar los resultados es necesario aclarar cual es la interpretación de los tipos de distancias o métricas implementadas:

El primero de la componentes, C_i, identifica el tipo de métrica. En función de la métrica seleccionada, el segundo de los campos o componentes representa el valor de $threshold_1$, $threshold_2$ o el valor de w_1 en los casos de las métricas C_1, C_2 y C_3 respectivamente. Finalmente, la última de las componentes representa el valor de w_1 en el caso únicamente de las métricas C_1 y C_2.

A modo de ejemplo se especifica el significado de la distancias coronadas como 1ª, 8ª y 10ª de la categoría cambio de ángulo:

C_1 | 0,7 | 0,66 $\longrightarrow$ Métrica *distancia umbralizada*, $threshold_1 = 0{,}7$, $w_1 = 0{,}66 \Rightarrow w_2 = 0{,}34$.

C_3 | 0 $\longrightarrow$ Métrica *distancia mínima simple*, $w_1 = 0 \Rightarrow w_2 = 1$.

C_2 | 0,5 | 0,8 $\longrightarrow$ Métrica *distancia basada en el vecino más próximo*, $threshold_2 = 0{,}5$, $w_1 = 0{,}8 \Rightarrow w_2 = 0{,}2$.

A partir de los resultados obtenidos se pueden elaborar diferentes conclusiones acerca de la optimización del descriptor.

El rango de la diferencia en el rendimiento del descriptor en las distintas categorías bajo análisis así como en una situación global es de aproximadamente 10 puntos, lo que significa que esta etapa de optimización resulta muy útil para valorar e identificar la mejor de las situaciones en las que el descriptor SIFT puede funcionar.

La mejor de las métricas implementadas resulta ser la denominada *distancia umbralizada* descrita mediante la Definición 2. Este primer puesto y varios más de los siguientes se ven representados por diferentes configuraciones de los demás parámetros pero siempre referidas a la métrica mencionada. Las otras 2 métricas se establecen en la mitad de la tabla de rendimientos alternándose según en que categoría.

Dentro de la métrica con mejor rendimiento, C_1, existen diferentes configuraciones de los demás parámetros $threshold_1$ y w_1, que alternan entre los primeros puestos, recordando al igual que en las optimizaciones de los anteriores descriptores, que la decisión final en cuanto a la elección de la métrica seleccionada se basa en una situación global mas representativa. Por este motivo se evita realizar aquí un análisis pormenorizado en cada una de las categorías y se identifica a la métrica $C_1 \mid threshold_1 = 0{,}5 \mid w_1 = 0{,}6$ como la mejor de las combinaciones con un rendimiento de 1 punto superior respecto de su más inmediata perseguidora.

Tipo de escena - Ranking		1º	2º	3º	4º	5º	6º	7º	8º	9º
Cambio de ángulo	Tipo distancia	$C_1\|0,7\|0,66$	$C_1\|0,5\|0,73$	$C_1\|0,8\|0,66$	$C_1\|0,5\|0,66$	$C_1\|0,5\|0,60$	$C_1\|0,3\|0,66$	$C_1\|0,5\|0,5$	$C_3\|0$	$C_3\|0,1$
	Valor	0.9874	0.9861	0.9849	0.9848	0.9834	0.9779	0.9667	0.965	0.9543
Cambio Iluminación	Tipo distancia	$C_1\|0,5\|0,6$	$C_2\|0,5\|0,8$	$C_1\|0,8\|0,66$	$C_1\|0,7\|0,66$	$C_2\|0,7\|0,7$	$C_1\|0,5\|0,66$	$C_2\|0,5\|0,6$	$C_2\|0,5\|0,7$	$C_3\|0$
	Valor	0.9945	0.9939	0.9925	0.9924	0.9905	0.9897	0.9882	0.9877	0.9872
Escena de Movimiento	Tipo distancia	$C_1\|0,5\|0,6$	$C_1\|0,3\|0,66$	$C_1\|0,5\|0,66$	$C_2\|0,5\|0,8$	$C_1\|0,5\|0,5$	$C_1\|0,7\|0,66$	$C_1\|0,5\|0,73$	$C_1\|0,8\|0,63$	$C_2\|0,5\|0,75$
	Valor	0.7764	0.7597	0.7412	0.7232	0.7099	0.6975	0.6842	0.6711	0.6674
Variación de Zoom	Tipo distancia	$C_1\|0,5\|0,43$	$C_2\|0,5\|0,8$	$C_1\|0,5\|0,5$	$C_1\|0,5\|0,6$	$C_2\|0,5\|0,75$	$C_3\|0$	$C_3\|0,1$	$C_1\|0,3\|0,66$	$C_2\|0,7\|0,7$
	Valor	1	1	0.9999	0.9998	0.9998	0.9992	0.9989	0.9987	0.9986
Escena Global	Tipo distancia	$C_1\|0,5\|0,6$	$C_1\|0,3\|0,66$	$C_1\|0,5\|0,66$	$C_2\|0,5\|0,8$	$C_1\|0,5\|0,5$	$C_1\|0,7\|0,66$	$C_1\|0,5\|0,73$	$C_3\|0$	$C_1\|0,8\|0,66$
	Valor	0.9385	0.9298	0.9264	0.9167	0.9149	0.9104	0.9033	0.9003	0.8984

10º	11º	12º	13º	14º	15º	16º	17º	18º	19º	20º
$C_2\|0,5\|0,8$	$C_1\|0,5\|0,43$	$C_2\|0,5\|0,75$	$C_2\|0,7\|0,7$	$C_3\|0,3$	$C_2\|0,5\|0,7$	$C_1\|0,5\|0,33$	$C_2\|0,3\|0,7$	$C_2\|0,5\|0,6$	$C_3\|0,5$	$C_2\|0,5\|0,65$
0.9498	0.9418	0.9393	0.9332	0.9115	0.9082	0.9042	0.8913	0.8858	0.8839	0.8821
$C_2\|0,5\|0,75$	$C_2\|0,5\|0,65$	$C_1\|0,5\|0,5$	$C_2\|0,3\|0,7$	$C_1\|0,3\|0,66$	$C_1\|0,5\|0,73$	$C_3\|0,1$	$C_1\|0,5\|0,43$	$C_1\|0,5\|0,33$	$C_3\|0,3$	$C_3\|0,5$
0.9867	0.9838	0.9833	0.9829	0.9828	0.9784	0.9748	0.9705	0.933	0.8707	0.3697
$C_1\|0,5\|0,43$	$C_3\|0$	$C_2\|0,5\|0,6$	$C_2\|0,7\|0,7$	$C_3\|0,1$	$C_2\|0,5\|0,7$	$C_2\|0,5\|0,65$	$C_2\|0,3\|0,7$	$C_1\|0,5\|0,33$	$C_3\|0,3$	$C_3\|0,5$
0.662	0.6496	0.6344	0.633	0.6043	0.5952	0.5787	0.5741	0.5692	0.4695	0.2449
$C_2\|0,5\|0,7$	$C_2\|0,3\|0,7$	$C_1\|0,5\|0,33$	$C_1\|0,5\|0,66$	$C_2\|0,5\|0,6$	$C_2\|0,5\|0,65$	$C_3\|0,3$	$C_1\|0,5\|0,73$	$C_1\|0,7\|0,66$	$C_1\|0,8\|0,66$	$C_3\|0,5$
0.9961	0.9929	0.9925	0.9898	0.9886	0.9883	0.9875	0.9645	0.9643	0.9449	0.4528
$C_2\|0,5\|0,75$	$C_1\|0,5\|0,43$	$C_2\|0,7\|0,7$	$C_3\|0,1$	$C_2\|0,5\|0,6$	$C_2\|0,5\|0,7$	$C_2\|0,3\|0,7$	$C_2\|0,5\|0,65$	$C_1\|0,5\|0,33$	$C_3\|0,3$	$C_3\|0,5$
0.8983	0.8936	0.8888	0.8831	0.8742	0.8718	0.8603	0.8582	0.8497	0.8098	0.4878

Tabla 4: Optimización descriptor SIFT

4.4.5. SURF

Al igual que en el caso del descriptor SIFT, el código del algoritmo que representa el comportamiento del descriptor SURF utilizado en este proyecto ha sido descargado a través la propia página web del autor[14].

El descriptor SURF comparte muchas características con el descriptor SIFT y por este motivo la optimización del mismo es compartida por el descriptor SIFT, exceptuando claro está, los resultados obtenidos. Todas las aportaciones y mejoras implementadas en el descriptor SIFT se encuentran en este apartado, coincidiendo tanto los valores de los parámetros como las métricas utilizadas. Por este motivo en esta sección tan sólo se identifican las peculiaridades y parámetros establecidos en base a la descripción teórica de la sección 3.5 y se atienden las aportaciones o diferencias que puedan existir en el código del algoritmo respecto del modelo teórico.

- $nOctaves = 4 \longrightarrow$ identifica el número de octavas que se utilizan en la búsqueda de puntos clave.
- $nOctaveLayers = 2 \longrightarrow$ es el valor por defecto del número de filtros o capas dentro de cada octava.
- $extended = 0 \longrightarrow$ valor por defecto que identifica la longitud del vector de características de 64 componentes; en el caso de establecerse como 1, el vector pasaría a estar formado por 128 componentes, al igual que en el caso del descriptor SIFT.

Los resultados obtenidos en esta etapa de optimización del descriptor SURF son recogidos por la Tabla 5. Dado que las métricas, parámetros y valores de los parámetros utilizados en la optimización del descriptor SURF son los mismos que para el descriptor SIFT, la interpretación de los resultados de la tabla se hace de la misma manera.

A tenor de los resultados mostrados, el rango de la diferencia entre los distintos rendimientos obtenidos por las diferentes métricas es mayor en el caso del descriptor SURF que en el del descriptor SIFT, alrededor de 14 puntos o 14 %. También resulta destacable que el rendimiento del descriptor SIFT en general supera en aproximadamente un 5 % el rendimiento mostrado por el descriptor SURF.

Sin embargo, la naturaleza de los resultados es compartida por ambos descriptores, siendo la métrica *distancia umbralizada* la que mayor rendimiento alcanza tanto en la mayor parte de las situaciones analizadas como en el caso de una situación global, aunque con diferentes valores de los parámetros.

Finalmente consta en los resultados que la mejor combinación posible de métricas y parámetros es la representada por: C_1 | 0,5 | 0,33.

[14] http://www.vision.ee.ethz.ch/~surf/download.html

Tipo de escena - Ranking		1º	2º	3º	4º	5º	6º	7º	8º	9º
Cambio de ángulo	Tipo distancia	C_1|0,5|0,33	C_1|0,3|0,66	C_1|0,5|0,5	C_1|0,5|0,43	C_1|0,5|0,60	C_1|0,5|0,66	C_1|0,5|0,73	C_1|0,7|0,66	C_3|0,1
	Valor	0.9925	0.9894	0.989	0.9884	0.9872	0.9871	0.9871	0.9634	0.9316
Cambio Iluminación	Tipo distancia	C_1|0,5|0,33	C_1|0,5|0,43	C_2|0,5|0,8	C_1|0,3|0,66	C_1|0,5|0,6	C_1|0,5|0,5	C_1|0,5|0,73	C_1|0,5|0,66	C_1|0,7|0,66
	Valor	0.9755	0.9681	0.9668	0.9633	0.9626	0.9626	0.9625	0.9625	0.9569
Escena de Movimiento	Tipo distancia	C_1|0,5|0,33	C_1|0,5|0,5	C_1|0,3|0,66	C_2|0,5|0,6	C_1|0,5|0,73	C_1|0,5|0,66	C_1|0,5|0,43	C_2|0,5|0,8	C_1|0,7|0,66
	Valor	0.6994	0.6973	0.6961	0.6934	0.6928	0.6926	0.6809	0.6702	0.6442
Variación de Zoom	Tipo distancia	C_2|0,5|0,8	C_2|0,5|0,75	C_2|0,7|0,7	C_1|0,5|0,6	C_2|0,5|0,7	C_3|0	C_2|0,5|0,65	C_2|0,3|0,7	C_3|0,3
	Valor	0.9951	0.995	0.9911	0.9891	0.9795	0.9772	0.9719	0.9696	0.9687
Escena Global	Tipo distancia	C_1|0,5|0,33	C_1|0,3|0,66	C_1|0,5|0,5	C_1|0,5|0,43	C_1|0,5|0,6	C_1|0,5|0,73	C_1|0,5|0,66	C_2|0,5|0,8	C_2|0,5|0,75
	Valor	0.9082	0.8982	0.8919	0.8908	0.8907	0.8905	0.8904	0.8884	0.871

10º	11º	12º	13º	14º	15º	16º	17º	18º	19º	20º
C_2|0,8|0,66	C_2|0,5|0,8	C_3|0	C_2|0,5|0,75	C_3|0,3	C_3|0,5	C_2|0,7|0,7	C_2|0,5|0,7	C_2|0,5|0,6	C_2|0,5|0,65	C_2|0,3|0,7
0.9285	0.9217	0.912	0.9083	0.9069	0.8959	0.88	0.8484	0.8427	0.8328	0.8254
C_2|0,5|0,75	C_1|0,8|0,66	C_2|0,7|0,7	C_2|0,5|0,7	C_3|0	C_2|0,5|0,6	C_2|0,3|0,7	C_2|0,5|0,65	C_3|0,1	C_3|0,3	C_3|0,5
0.9553	0.9538	0.9521	0.9468	0.9456	0.9417	0.9362	0.9322	0.9309	0.8921	0.7938
C_2|0,5|0,6	C_2|0,5|0,75	C_1|0,8|0,66	C_2|0,7|0,7	C_2|0,5|0,65	C_3|0,1	C_3|0	C_2|0,5|0,7	C_2|0,3|0,7	C_3|0,3	C_3|0,5
0.6299	0.6253	0.6016	0.5853	0.5688	0.5555	0.5474	0.5347	0.5049	0.476	0.3962
C_1|0,5|0,33	C_3|0,1	C_1|0,3|0,66	C_3|0,5	C_2|0,5|0,43	C_1|0,5|0,66	C_1|0,5|0,6	C_1|0,5|0,73	C_1|0,5|0,5	C_1|0,7|0,66	C_1|0,8|0,66
0.9655	0.9624	0.9441	0.9344	0.9256	0.9195	0.9195	0.9194	0.9189	0.8776	0.8485
C_1|0,7|0,66	C_2|0,7|0,7	C_2|0,5|0,6	C_3|0	C_3|0,1	C_1|0,8|0,66	C_2|0,5|0,7	C_2|0,5|0,65	C_3|0,3	C_2|0,3|0,7	C_3|0,5
0.8605	0.8521	0.8509	0.8455	0.8451	0.8331	0.8274	0.8264	0.8109	0.809	0.7551

Tabla 5: Optimización descriptor SURF

4.5. Comparación de Descriptores: Análisis inter-descriptor

Una vez los distintos descriptores han sido optimizados con sus respectivas parametrizaciones, en esta etapa se lleva a cabo un estudio comparativo entre todos ellos para identificar que descriptor y para que situaciones resultan más apropiados.

Para llevar a cabo este estudio, se ha utilizado un nuevo conjunto de imágenes denominado test data set mencionado en la sección 4.2, y que está compuesto por el conjunto completo de toda la base de datos de imágenes. Sin embargo, para realizar el estudio comparativo se han seleccionado 400 imágenes querys y a cada una de ellas se le asigna un conjunto de 200 imágenes totalmente aleatorias del conjunto completo de la base de datos. Al igual que en el conjunto optimization data set, los grupos asignados a cada una de las imágenes query contienen las imágenes relacionadas con la escena de la imagen query y el resto hasta las 200 imágenes se eligen ahora si, de forma aleatoria del total de la base de datos.

Los resultados se muestran mediante la Tabla 6 y también de forma gráfica mediante los diferentes gráficos de las curvas de Precision-Recall (PR) de cada uno de los descriptores presentados en cada una de los apartados en los que esta sección se encuentra dividida.

El análisis de los resultados se realiza diferenciando cada una de las 4 categorías analizadas para finalmente analizar un escenario global ya mencionado en la etapa de optimización.

4.5.1. Cambios de ángulo

Las escenas con cambios de ángulo se caracterizan por la aparición de nuevas regiones o por la ocultación de otras anteriormente presentes en la imagen original, y por lo tanto una gran variedad de la información dependiendo del ángulo de variación de la escena. Sin embargo, y en base a las imágenes de esta naturaleza contenidas en la base de datos, estas variaciones no suelen afectar a la composición de colores presentes en la imagen original.

Teniendo en cuenta todo lo anterior, parecen más razonables los resultados recogidos en la tabla con respecto a esta categoría. En ella se puede comprobar como el rendimiento de los descriptores que guardan una relación más directa con los colores de la imagen, presentan un mayor rendimiento que los descriptores basados en la posición de los puntos clave como son los descriptores SIFT y SURF.

El histograma HSV obtiene el mejor de los rendimientos en cuanto a la detección de imágenes afectadas por cambios de ángulo, con un margen de al menos un 3 % sobre el resto de los descriptores globales, y con casi un 7 % de rendimiento positivo con respecto a los descriptores locales. Tras el histograma HSV se encuentran los descriptores Correlograma, histograma RGB y Color Layout con un rendimiento muy similar. Finalmente, los descriptores SIFT y SURF

Tipo de escena - Ranking	1º	2º	3º	4º	5º	6º
Cambio de ángulo	Descriptor	Histograma HSV	Correlograma	Histograma RGB	Color Layout	SIFT	SURF
	Valor	0.9748	0.9428	0.9426	0.9376	0.9268	0.9
Cambio Iluminación	Descriptor	SIFT	SURF	Histograma HSV	Histograma RGB	Color Layout	Correlograma
	Valor	0.9506	0.9258	0.7114	0.6389	0.5434	0.4326
Escena de Movimiento	Descriptor	Histograma HSV	SIFT	SURF	Histograma RGB	Color Layout	Correlograma
	Valor	0.8905	0.847	0.8397	0.8294	0.7848	0.7613
Variación de Zoom	Descriptor	SURF	SIFT	Histograma HSV	Color Layout	Histograma RGB	Correlograma
	Valor	0.9682	0.9652	0,9576	0.8628	0.8378	0.7415
Escena Global	Descriptor	SIFT	SURF	Histograma HSV	Histograma RGB	Color Layout	Correlograma
	Valor	0.9112	0.9084	0.8835	0.8122	0,7819	0.7195

Tabla 6: Comparación de descriptores

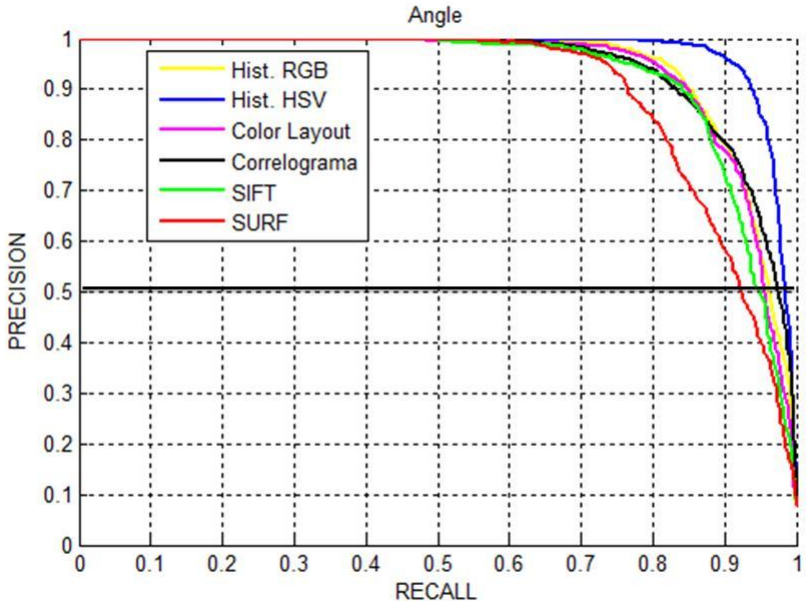

Figura 4.10: Comparación descriptores cambio de ángulo

son penalizados con un rendimiento inferior respecto de los anteriores. Las imágenes afectadas por cambios de ángulo presentan oclusión de algunas regiones y la aparición de otras nuevas como simple consecuencia del giro del ángulo de visión. Esta peculiaridad afecta en mayor medida a los descriptores locales SIFT y SURF debido a que el número de puntos emparejados desciende como consecuencia de las peculiaridades mencionadas. Sin embargo la característica del color puede verse menos afectada de forma global ante estas peculiaridades, cuando las nuevas regiones y las que han quedado ocultas se componen de colores similares. Incluso en estas situaciones, los descriptores SIFT y SURF mantienen en un rendimiento del 90 % con respecto a la detección de imágenes similares y afectadas por los cambios de ángulo pertenecientes a la misma escena.

La Figura 4.10 representa el comportamiento de los descriptores mediante las curvas PR. Adicionalmente se puede utilizar estas representaciones para realizar un juicio rápido sobre el rendimiento de los descriptores mediante el valor de Recall para un valor de Precision de 0.5, $Recall\ |_{Precision=0,5}$. De esta manera se observa como el valor de Recall va desde 0,92 para el caso del descriptor SIFT hasta el 0,98 en el caso del histograma HSV.

4.5.2. Cambios de iluminación

Al contrario que los cambios de ángulo anteriores, los cambios de iluminación afectan directamente al valor de los pixels, modificando de esta manera lo colores

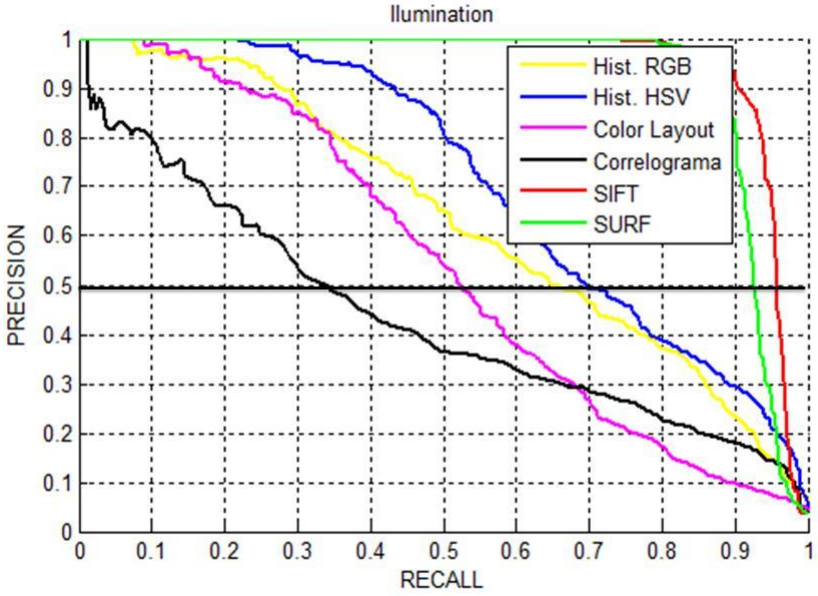

Figura 4.11: Comparación descriptores cambio de iluminación

presentes en la imagen, sin variar apenas la morfología de la composición.

Por este motivo, todos aquellos descriptores que están basados directamente en el color de la imagen, registran un rendimiento muy pobre ante este tipo de situaciones. Con respecto a los datos presentes en la Tabla 6 se aprecia de forma muy clara como el rendimiento de los descriptores SIFT y SURF, con resultados del 95 % y 92 % de éxito respectivamente en la detección de este tipo de imágenes relacionadas, resulta inútilmente comparado con el rendimiento del resto de descriptores elegidos, donde en el mejor de los casos se alcanza un escaso 70 % de éxito para el descriptor histograma HSV y decayendo el rendimiento en un 10 % adicional y aditivo para cada uno de los descriptores restantes.

Como prueba adicional se muestran en la Figura 4.11 la curvas PR correspondientes a la actuación de los descriptores ante imágenes con cambios de iluminación presentes. Finalmente, los valores de $Recall\,|_{Precision=0,5}$ para cada uno de los descriptores son, ordenados según la importancia de los resultados: (0,96, 0,93, 0,7, 0,66, 0,52, 0,33) para los descriptores SIFT, SURF, histograma HSV, histograma RGB, Color Layout, Correlograma.

4.5.3. Escenas con movimiento de objetos

Las imágenes pertenecientes a escenas donde se intenta analizar el movimiento de diferentes objetos y regiones representan una gran dificultad debido a que ni siguen un mismo patrón de movimiento todos los objetos presentes en la im-

agen ni tampoco todos las imágenes afectadas por estas transformaciones comparten la naturaleza de un tipo de movimientos en particular. Como ejemplo se puede mencionar la dificultad que entraña relacionar las diferentes imágenes que representan los distintos movimientos de las personas recogidos en la filmación de una escena de una conversación en un ambiente público.

Los resultados obtenidos tras la comparación de los descriptores respecto de este tipo de escenas, recogidos en la Tabla 6, muestran las dificultades mencionadas registrando un rendimiento en el mejor de los casos por debajo del 90 % de efectividad respecto de la detección de imágenes pertenecientes a escenas cuasi-duplicadas. Este empobrecimiento en el rendimiento supone una disminución de más del 6 % respecto del mejor de los rendimientos en las otras 3 categorías analizadas.

Más concretamente, el mejor de los descriptores vuelve a ser de nuevo el histograma HSV con un margen positivo respecto del peor de los descriptores, nuevamente el correlograma, de más de un 13 %. A continuación se encuentran en la escala de rendimientos el descriptor SIFT primero y después SURF con diferencias que rondan el 5 % y 6 % respectivamente. Estas diferencias pueden deberse a la mayor dificultad de los descriptores locales de establecer correspondencias cuando están presentes movimientos en las imágenes que pueden suponer oclusiones de algunas regiones o la aparición de otras nuevas. Al igual que en la categoría previa de imágenes afectadas por cambios de iluminación, los descriptores Color Layout y Correlograma parecen descolgarse de las prestaciones ofrecidas por el resto de los descriptores con rendimientos que no superan el 80 %.

En la gráfica de la Figura 4.12 se pueden apreciar las diferencias comentadas mediante la representación de las curvas PR de los distintos descriptores así como mediante los valores de *Recall* $|_{Precision=0,5}$ correspondientes, más concretamente: (0,905, 0,86, 0,846, 0,844, 0,836, 0,8) para los descriptores histograma HSV, SIFT, SURF, histograma RGB, Color Layout y Correlograma respectivamente.

4.5.4. Variaciones de zoom

Los resultados obtenidos para la detección de imágenes similares respecto la presencia de variaciones de zoom entre ellas presentes en la Tabla 6 muestran la clasificación de los rendimientos en dos grupos diferenciados. El primero de los grupos formado por los descriptores SIFT, SURF e histograma HSV alcanza rendimientos en torno al 95 % de éxito a la hora de la detección de este tipo de peculiaridades en las imágenes. El segundo de los grupos está formado por los tres descriptores restantes, histograma RGB, Color Layout y el correlograma. Entre ambos grupos existe un margen de entre el 10 % y el 25 % en el rendimiento respecto del histograma RGB y el peor de todos ellos, el correlograma.

La separación entre ambos grupos pone a su vez de manifiesto las mejores cualidades de los descriptores locales frente a los globales para la detección

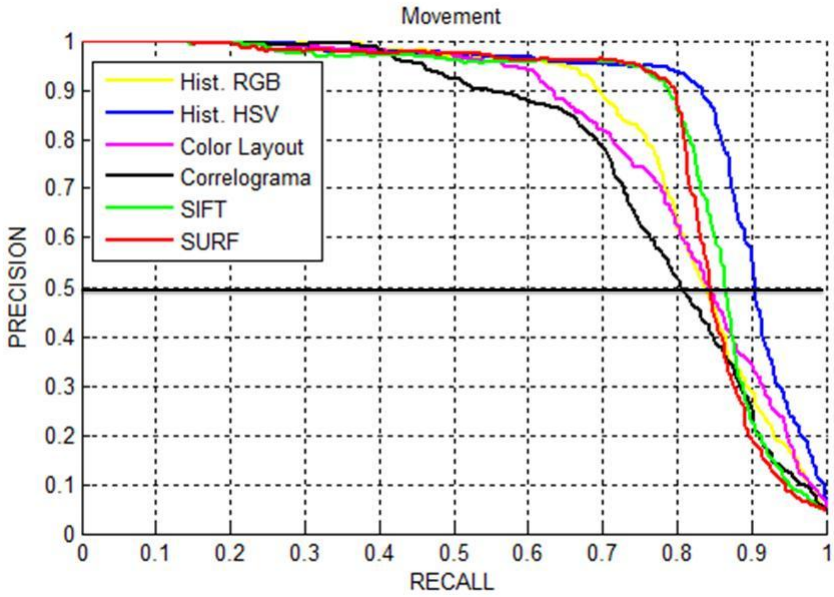

Figura 4.12: Comparación descriptores en escenas con movimiento de los objetos que la componen

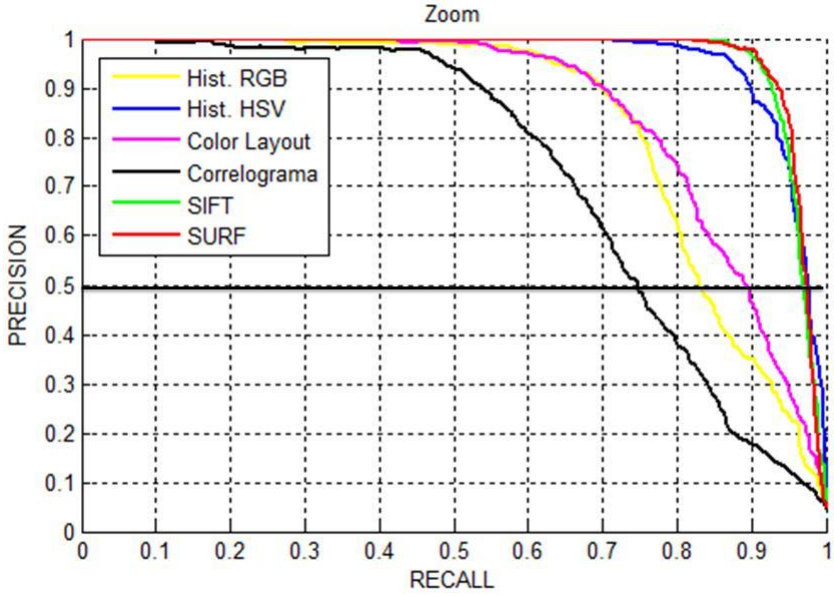

Figura 4.13: Comparación descriptores variación de zoom

de imágenes pertenecientes a una misma escena y relacionadas mediante la variación de zoom. Esta ventaja reside en el hecho de que al comparar dos imágenes con una variación de zoom entre ellas, si bien las relaciones locales entre los valores de los pixels no sufren apenas variaciones, y por tanto los detectores como SIFT y SURF son capaces de detectar las correspondencias mediante la aplicación de diferentes escalas (ver secciones 3.4 y 3.5), la características globales de la imagen varían en mayor medida cuanto mayor es el índice de zoom entre ambas imágenes. Sin embargo, cabe destacar el rendimiento del histograma HSV incluyéndose dentro del primero de los grupos, aún siendo de naturaleza global.

Finalmente son representadas las curvas PR de los descriptores en la Figura 4.13. Para disponer de el mismo conjunto de datos que en los casos analizados anteriores, se presentan a continuación los valores de $Recall\ |_{Precision=0,5}$: (0,977, 0,974, 0,97, 0,895, 0,83, 0,747) respecto de los descriptores SIFT, SURF, histograma HSV, histograma RGB, Color Layout y Correlograma.

4.5.5. Análisis de un escenario global

Tras la exposición de los resultados obtenidos en cada una de las categorías de análisis del proyecto, y manteniendo el argumento presentado en la sección 4.4 sobre la optimización de los descriptores, se presenta en esta sección las valoraciones y análisis de una situación que representa un escenario global teniendo en cuenta la diferentes categorías.

Como aval de las conclusiones aquí expuestas se hace referencia a los resultados obtenidos en la Tabla 4.5 para un escenario global y que han sido calculados como una media de los resultados obtenidos en las 4 categorías restantes. Como primera apreciación cabe destacar la creación de tres grupos representativos de los de rendimientos de los descriptores. Al igual que para el caso de las imágenes afectadas por variaciones de zoom, el primero de los grupos está formado por los descriptores SIFT, SURF e histograma HSV. Este grupo representa a los descriptores que califican con un rendimiento medio alrededor del 90 %. El segundo de los grupos lo conforman el histograma RGB y el Color Layout Descriptor con rendimientos en torno al 80 %. Como único representante del tercer y último grupo se encuentra el correlograma con un rendimiento medio de apenas el 70 %.

Si bien es cierto que existen notables diferencias en cuanto a los resultados del rendimiento por parte de ambos histogramas de color utilizados en este proyecto, histogramas RGB y HSV con una diferencia alrededor del 8 %, resultaría injusta la elaboración de una conclusión que diera al histograma HSV como claro ganador en detrimento del histograma RGB. La diferencia entre el número de bins utilizados en uno y otro caso ($32^3 = 32768$ y $16x4x4 = 256$ representaciones de colores diferentes para el caso del histograma RGB y HSV respectivamente) mantiene la imposibilidad de una comparación entre ambos

de forma justa. Por este motivo los resultados aquí mostrados hacen referencia al caso particular implementado, obedeciendo a lo establecido en la sección 4.4, quedando como trabajo futuro la comprobación o refuta miento de la apreciación aquí mostrada mediante la variación y equiparación en el numero de bins de ambos histogramas.

Respecto de los descriptores SIFT y SURF se puede mencionar que su comportamiento es bastante parecido en todas y cada una de las categorías analizadas con diferencias mínimas entre ambos, aunque en la mayoría de la situaciones el descriptor SURF se muestra como superior.

Mediante el resumen de los distintos comportamientos de los descriptores se llega a la conclusión de que son los descriptores locales, en detrimento de los globales, los que alcanzan un mayor rendimiento en cuanto a la tarea de detección de imágenes cuasi-duplicadas y pertenecientes a una misma escena se refiere. Cabe destacar la sorpresa en el rendimiento mostrado por el descriptor histograma HSV a pesar de ser un descriptor de características globales.

Como nota a tener en cuenta se menciona que los resultados obtenidos no dejan de ser dependientes de la morfología y del contenido de la base de datos utilizada en este proyecto, y si bien las tendencias o relatividad entre los diferentes rendimientos de los descriptores pueden mantenerse, el valor absoluto de los mismos se presume reduciría en el caso de contar tanto con una mayor cantidad de imágenes en la base de datos, con el aumento de los grupos de comparación, como con la introducción de más variedad en las categorías representadas.

En el Anexo II se incluyen ejemplos de las imágenes resultado de las comparaciones realizadas por los distintos descriptores.

4.6. Combinación de Descriptores

Una de las motivaciones de este proyecto expresada en la sección 1.1 ha sido la de realizar un estudio comparativo de los descriptores con respecto al marco de referencia establecido mediante la morfología de la base de datos. Se han realizado diferentes estudios sobre la comparación individual de los descriptores si bien la temática o marco de referencia son diferentes al mostrado en este proyecto. Sin embargo en esta sección se va un paso más allá y se evalúa el comportamiento de los descriptores de manera combinada, es decir, fusionando los descriptores mediante parejas. De esta manera se pretende mejorar los resultados anteriores y aprovechar las posibles asociaciones positivas que puedan surgir respecto a la detección de este tipo de imágenes contribuyendo así al avance del estado del arte.

La combinación de los descriptores se realiza mediante la ponderación de los resultados obtenidos por cada uno de los descriptores. Dado que los resultados de la distancia euclídea obtenidos mediante la aplicación individual han sido normalizados, la ponderación de las mismas mediante un valor también normalizado contribuye a la obtención de valores de distancia nuevamente normalizados y libres de cualquier interpretación errónea. La ecuación 4.17 representa la nueva métrica de combinación de los descriptores implicados obteniendo así una nueva distancia euclídea entre la imagen original y la imagen comparada:

$$Dist = [w(d_{Descriptor1}) + (1 - w)\,(d_{Descriptor2})] \tag{4.17}$$

donde $w \in [0, 1]$ es el factor que representa el peso de cada una de las componentes y $d_{Descriptor1}$, $d_{Descriptor2}$ representan el valor de la distancia entre las imágenes o *score* respecto de los descriptores que componen la combinación. El valor de w ha sido parametrizado para cubrir así más posibilidades y obtener unas conclusiones mejor fundamentadas. En concreto los valores han sido: $w =$ 0,1, 0,2, ..., 0,9. Además de esta métrica ponderada se ha implementado la media geométrica representada mediante la ecuación 4.18:

$$Dist = \sqrt{d_{Descriptor1} * d_{Descriptor2}} \tag{4.18}$$

Para cada una de las métricas mencionadas el número de combinaciones con respecto a las diferentes componentes viene determinado por las combinación de dos descriptores sin repetición, por lo tanto, $\frac{(n(n-1))}{2} = \frac{6*5}{2} = 15$ combinaciones diferentes. Dado que el número de descriptores de este proyecto no es muy elevado, se pueden cubrir todas las combinaciones posibles sin realizar un estudio sobre la correlación entre los diferentes descriptores necesario en caso de un numero mayor de descriptores.

Teniendo en cuenta tanto la parametrización de las métricas combinadas como el número de combinaciones diferentes con respecto a las componentes de las métricas, se obtienen como resultado $15_{combinaciones} * 10_{métricas} = 150$ combinaciones diferentes. Ante la imposibilidad de mostrar todos los resultados, que han sido obtenidos, organizados en una tabla se presentan tan sólo los

referentes a la métrica que mejor rendimiento reporta. Estos resultados son presentados en la Tabla 7 y pertenecen a la métrica ponderada con valor $w = 0{,}2$, lo que significa que la contribución del primero de los descriptores de cada pareja es tan sólo del 20 % sobre el resultado final.

Los resultados aquí mostrados han de ser analizados en comparación con los de la Tabla 6 que hacen referencia al rendimiento de los descriptores de manera individual. Según esto, y respecto de las distintas categorías pueden hacerse distintas apreciaciones.

Como primera aproximación se llevan a cabo las comparaciones entre la mejor combinación y el mejor de los descriptores en cada una de las categorías.

- Cambio de ángulo: La combinación HSV-SURF obtiene un rendimiento ligeramente superior que el descriptor HSV de forma individual.
- Cambio de iluminación: En este caso, el descriptor SIFT en su faceta individual se ve superado de nuevo por una combinación de descriptores a la que se une en este caso el descriptor HSV.
- Movimiento de objetos en la imagen: Sucede lo mismo que en el caso del cambio de ángulo, donde la combinación HSV-SURF está por encima del descriptor HSV.
- Variación de Zoom: En este caso, si bien la combinación de descriptores supera el rendimiento del descriptor aislado, el porcentaje es netamente superior y alrededor de una 3 %, lo que destaca la aportación del descripto HSV al rendimiento ya alcanzado por el descriptor SURF.
- Escenario Global: Es en esté análisis donde se merecen las razones del estudio combinativo, alcanzando la pareja de descriptores HSV-SURF una mejora sustancial de más del 5 % respecto del descriptor individual SIFT.

Como segunda estrategia de comparación se evalúa el rendimiento de las combinaciones en todas las categorías enfrentando tan sólo la pareja con mayor rendimiento global, HSV-SIFT, con el mejor de los descriptores individual, SIFT.

Si ya el rendimiento de ambos respecto de un escenario global ha sido analizado fallando a favor de la combinación de manera clara, todavía se ve más reforzada la combinación por el hecho de que en las situaciones de cambio de ángulo y movimiento en las imágenes, el rendimiento alcanzado por la combinación registra una mejora en la línea de la situación global, con una diferencia de más del 5 %.

Todavía resulta más fructífera la combinación de los descriptores HSV-SURF si se compara con los mismos de forma individual.

Teniendo en cuenta lo anterior, junto con el hecho de que la importancia del análisis se ve concretada en la categoría de un escenario global por ser más realista, se concluye que la combinación de los descriptores histograma HSV-SURF mejora los resultados obtenidos por el descriptor SIFT que a su vez representa el mejor de los rendimientos individual.

Tipo de escena - Ranking		1º	2º	3º	4º	5º	6º
Cambio de ángulo	Descriptor	HSV - SURF	HSV - SIFT	RGB - SURF	CL - HSV	RGB - SIFT	CRLG - HSV
	Valor	0.9882	0.9873	0.976	0.9759	0.9754	0.9751
Cambio Iluminación	Descriptor	HSV - SIFT	CRLG - SIFT	HSV - SURF	RGB - SIFT	SIFT - SURF	RGB - SURF
	Valor	0.9609	0.9537	0.9514	0.9434	0.9409	0.9316
Escena de Movimiento	Descriptor	HSV - SURF	HSV - SIFT	RGB - SIFT	CRLG - HSV	RGB - SURF	CL - HSV
	Valor	0.9056	0.9035	0.8925	0.891	0.8891	0.8885
Variación de Zoom	Descriptor	HSV - SIFT	HSV - SURF	SIFT - SURF	RGB - SIFT	CRLG - SURF	CRLG - SIFT
	Valor	0.9946	0.9927	0.9738	0.9716	0.9702	0.9702
Escena Global	Descriptor	HSV - SURF	HSV - SIFT	RGB - SIFT	RGB - SURF	CRLG - SIFT	SIFT - SURF
	Valor	0.9616	0.9595	0.9457	0.9409	0.9304	0.9217

7º	8º	9º	10º	11º	12º	13º	14º	15º
CL - RGB	HSV-RGB	CL - SURF	CL - SIFT	CL - CRLG	CRLG - RGB	CRLG - SIFT	SIFT - SURF	CRLG - SURF
0.9719	0.9608	0.9568	0.954	0.943	0.943	0.9406	0.9239	0.9155
CRLG - SURF	CL - SURF	CL - SIFT	CRLG - HSV	CL - HSV	CL - RGB	HSV-RGB	CRLG - RGB	CL - CRLG
0.93	0.773	0.7614	0.7109	0.6955	0.6942	0.6751	0.6384	0.5559
CL - RGB	CL - SURF	CL - SIFT	HSV-RGB	CRLG - SIFT	SIFT - SURF	CRLG - SURF	CRLG - RGB	CL - CRLG
0.888	0.8735	0.8674	0.8635	0.8571	0.8484	0.8468	0.8298	0.8285
RGB - SURF	CRLG - HSV	CL - HSV	CL - SURF	CL - RGB	CL - SIFT	HSV-RGB	CL - CRLG	CRLG - RGB
0.9668	0.9618	0.9453	0.9359	0.9294	0.927	0.8921	0.8713	0.8375
CRLG - SURF	CL - SURF	CRLG - HSV	CL - SIFT	CL - HSV	CL - RGB	HSV-RGB	CRLG - RGB	CL - CRLG
0.9156	0.8848	0.8847	0.8774	0.8763	0.8709	0.8479	0.8122	0.7997

Tabla 7: Combinación de descriptores

HSV → histograma HSV; RGB → histograma RGB; CL → color layout; CRLG → correlograma.

En el Anexo II se incluyen diferentes ejemplos de las imágenes resultado de las comparaciones realizadas por los distintos descriptores.

Finalmente se representa mediante las Figura 4.14la diferencia entre las curvas de PR respecto de la combinación HSV-SURF y el descriptor SIFT en cada una de las categorías de análisis, con el fin de resultar más claro y evidente la mejora en el rendimiento.

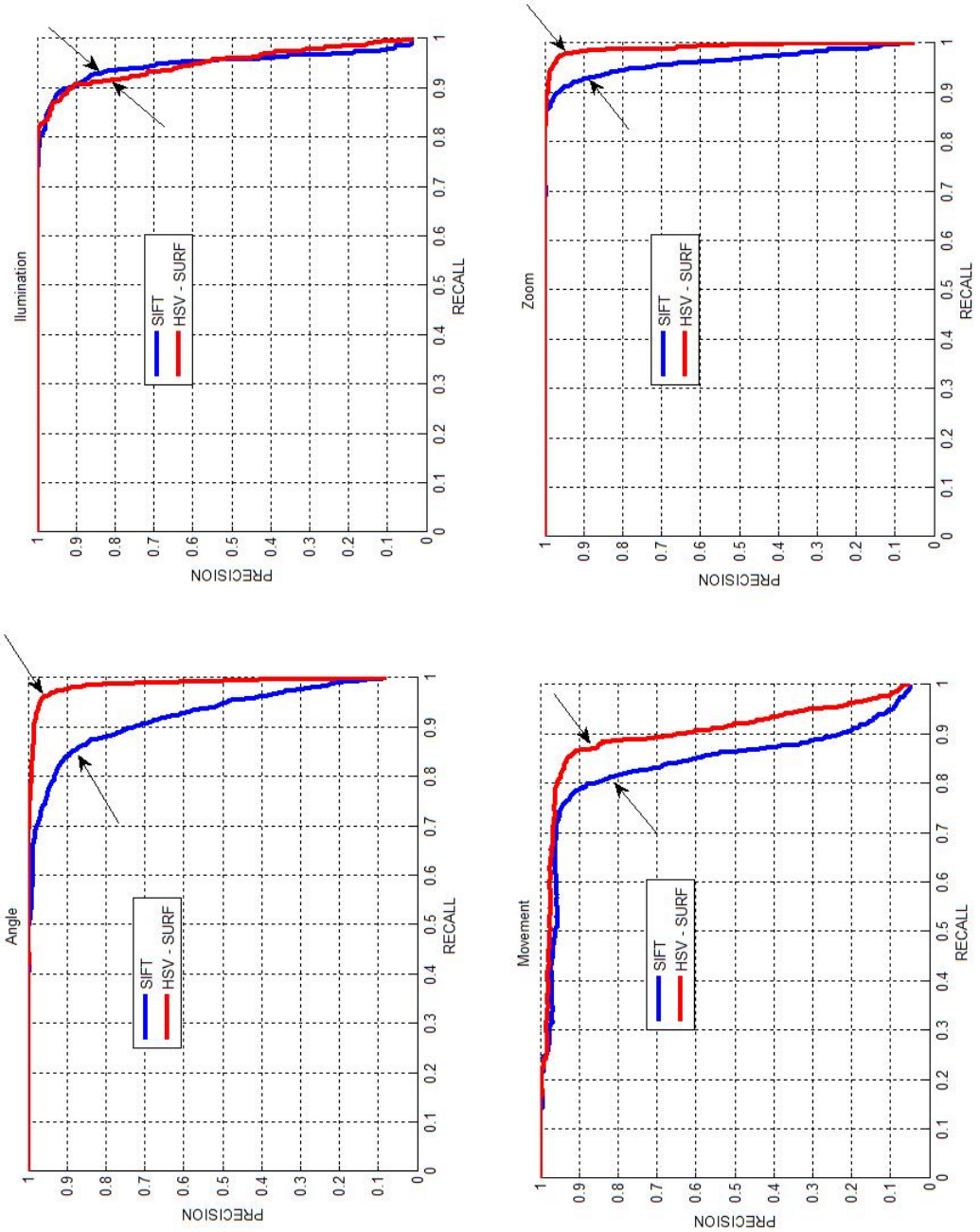

Figura 4.14: Diferencia entre descriptor SIFT y combinación de descriptores HSV-SURF

4.7. Coste Computacional

Uno de los aspectos más importantes a tener en cuenta cuando se realiza un estudio comparativo es el de establecer un marco de análisis que determine los puntos en común donde han de evaluarse los diferentes descriptores. Las conclusiones elaboradas a partir de los resultados serán más completas y precisas en función de la pluralidad y la independencia de los diferentes aspectos comparados. En concreto se puede dar el caso de que a pesar de obtener mejores resultados en cuanto al rendimiento, sea el coste computacional el que haga variar las decisiones y conclusiones elaboradas para la elección de uno u otro descriptor. Por esta razón en este proyecto se analiza, además del rendimiento que alcanzan los distintos descriptores en a la precisión de los resultados obtenidos en la detección de duplicados, el coste computacional de los mismos en las diferentes etapas de las que se compone el proceso.

El análisis del coste computacional de los descriptores se divide en diferentes apartados que tienen que ver con el tiempo empleado en la creación de los descriptores y en la tarea de comparación de las imágenes así como el tamaño o espacio requerido para almacenar los resultados vertidos por las dos etapas anteriores.

Las estimaciones y conclusiones elaboradas han de tener como referencia los resultados relativos al peor de los casos posibles. Por este motivo los datos mostrados en la Tabla 8 sobre los tiempos de creación, comparación y tamaño de los descriptores representan el peor de los resultados registrados durante la simulación de las distintas pruebas.

Los tiempos de creación y comparación han sido medidos mediante la programación de esta opción en el lenguaje Matlab® en el que se ha implementado el código de los distintos descriptores. El tamaño de los mismos también está sujeto a la representación que Matlab® realiza sobre los datos de tipo *double*. En concreto, se reservan 4 bytes de espacio para almacenar cada dato de tipo *double*, por lo que todos los datos presentes en la tabla están sujetos a esta sujetos a esta conversión. Es necesario dejar constancia de que tanto el tiempo como el espacio requerido pueden variar en función del lenguaje de programación elegido o la máquina en la que se lleve a cabo el proceso, sin embargo parece razonable aceptar que las relaciones expresadas mediante estos resultados han de mantenerse constantes con independencia de la elección tomada.

Tiempo de creación: Este aspecto adquiere una gran importancia cuando se trata de aplicaciones en tiempo real. En algunos casos esta tarea puede hacerse de manera *off-line,* evitando así este inconveniente, pero de no ser así y en relación a los datos mostrados en la tabla, esta etapa puede convertirse en un llamado cuello de botella dependiendo de la magnitud del conjunto de imágenes. Con respecto a los datos expuestos en la Tabla 8, existen dos grupos de descriptores bien diferenciados: i) Ambos histogramas RGB y HSV,

Descriptor	Tiempo Creación (ms)	Tiempo Comparación (ms)	Tamaño Descriptor
RGB Histogram	130 ms.	6 ms.	384 bytes // 1536 bytes
HSV Histogram	290 ms.	5,4 ms.	96 bytes // 384 bytes
Color Layout	60 ms.	2,5 ms.	768 bytes
Correlogram	9800 ms.	5,5 ms.	960 bytes ≈ 1 Kb
SIFT	7100 ms.	600 ms.	470 Kb
SURF	350 ms.	125 ms.	100 Kb

Tabla 8: Coste computacional de descriptores

Color Layout y SURF pertenecen al grupo donde el tiempo medio de creación varía entre decenas y pocos cientos de milisegundos; ii) el otro de los grupos en los que se incluyen el descriptor SIFT y el correlograma requieren un tiempo de creación de mas de 5 segundos por cada imagen, lo que supone un incremento de 20 veces más respecto del primero de los grupos. Esto implica que ambos descriptores del segundo de los grupos quedan reducidos a la aplicación en situaciones donde no exista la restricción de funcionamiento en tiempo real.

Tiempo de comparación: El tiempo requerido para realizar las distintas comparaciones entre las imágenes guarda una relación más estrecha con el tamaño del conjunto de datos que con la modalidad de la aplicación. En el caso que aquí se presenta existen 3 grupos diferenciados: i) El primero de los grupos esta representado por un tiempo de comparación de escasos milisegundos, y en el tienen cabida los 4 descriptores globales que se evalúan en este proyecto como son Color Layout, Correlograma, y los histogramas HSV y RGB; ii) En una escala superior se encuentra el descriptor SURF representado por un tiempo que no supera las 2 décimas de segundo por cada comparación entre dos imágenes; iii) En otra escala diferente se encuentra el descriptor SIFT donde el tiempo medio de comparación supera el $\frac{1}{2}$ segundo y resulta incomparable respecto del primero de los grupos i). Como resumen se puede interpretar que el descriptor SIFT requiere un tiempo de comparación por encima de 100 veces más respecto del grupo i), y más de 4 veces superior respecto del descriptor SURF.

Tamaño del descriptor: Con respecto al tamaño requerido para alojar los diferentes descriptores, se pueden discernir claramente 3 grupos nuevamente: i) el primero de ellos compuesto nuevamente por los descriptores globales, cuyo espacio medio requerido no supera los 2 Kb; ii) el segundo grupo al que pertenece el descriptor SURF y que representado con una escala de unas 50 veces más respecto del grupo i) requiere un espacio de 100 Kb; iii) nuevamente en el grupo más restrictivo se encuentra el descriptor SIFT donde con casi $\frac{1}{2}$ Mb representa 5 veces el espacio requerido por el descriptor SURF y con un orden de magnitud de 10^2 sobre el grupo i).

Tras el análisis del coste computacional y junto con las conclusiones presentadas en la sección 4.6 se pueden elaborar ahora de una manera más contundente las siguientes conclusiones:

- Dado que las combinaciones entre descriptores globales exclusivamente no obtienen un rendimiento superior con respecto al rendimiento alcanzado por los descriptores locales de forma individual, la mayor de las ventajas respecto al bajo coste que estos representan respecto de los descriptores locales, no es explotada de forma completa. Por este motivo han de relajarse las restricciones para descubrir que, si bien el descriptor SURF presenta un coste computacional medio, la combinación de descriptores HSV - SURF obtiene no sólo una mejora del rendimiento en relación a la capacidad de detección frente al mejor de los rendimientos individuales, el descriptor SIFT, sino también una mejora del coste computacional que supone la combinación respecto del descriptor SIFT aún de forma individual. Esto implica una mejora mayor y mejor fundamentada de la combinación de ambos descriptores sobre el descriptor SIFT individual.

- Además, existen combinaciones de descriptores globales, como por ejemplo el caso de CL - RGB, que si bien de forma individual alcanzan a lo sumo un rendimiento del 81 %, de forma combinada alcanzan un rendimiento notablemente mayor de hasta el 87 % con un incremento del coste computacional por parte de la combinación de ambos con respecto a los descriptores individualmente de todas formas despreciable respecto del aumento del rendimiento alcanzado, más de un 6 % (ver Tablas 6 y 7).

5. Conclusiones y trabajo futuro

En el presente Proyecto se expone un estudio comparativo de diferentes descriptores visuales compuesto por tres etapas diferenciadas: optimización, comparación y combinación de los descriptores seleccionados.

Previamente a las etapas mencionadas, existe un trabajo de documentación y estudio del estado del arte de los descriptores visuales que desemboca en la selección de los descriptores utilizados. A pesar de la dificultad de sugerir o dar respuesta a la pregunta sobre que descriptores son mejores dependiendo para que tareas [4, 12], esta selección está apoyada en los argumentos de variedad, usabilidad y representatividad para reunir a los descriptores locales SIFT y SURF así como a los descriptores globales Histograma de color RGB y HSV, Color Layout y Correlograma de color.

La evaluación del rendimiento aportado por cada uno de los descriptores está fundamentada sobre la creación de una base de datos de imágenes propia que representa las diferentes categorías del estudio relativas a cuatro tipos de transformaciones de imagen: cambios de ángulo, variaciones de zoom, cambios de iluminación y movimiento de los objetos presentes en la escena. Estas categorías forman parte del variado elenco de transformaciones que son objeto de estudio para tareas como la generación automática de resúmenes de vídeo, sistemas CBIR, etc.

La primera de las etapas somete a los distintos descriptores a diferentes optimizaciones. La importancia de esta etapa radica en la obtención de un marco de comparación sobre el que se edifican las presentaciones de los avances conseguidos así como el punto de encuentro entre las diferentes etapas sucesivas de comparación de los descriptores ya optimizados y de las diferentes combinaciones de los mismos. Los resultados obtenidos en esta primera etapa aportan claridad sobre las diferentes propuestas de mejora llevadas a cabo sobre cada uno de los descriptores contribuyendo de esta manera a la elección de los parámetros que resultan óptimos en cada caso.

Dentro de las distintas pruebas llevadas a cabo en esta etapa de optimización se han incluido distintas métricas para el cómputo del matching score entre dos imágenes así como la parametrización de otras. Merece especial mención la mejora propuesta sobre la eliminación de falsas correspondencias entre puntos clave mediante restricciones geométricas para los descriptores locales. El razonamiento detrás de esta mejora es simple: de tratarse de dos imágenes relacionadas mediante algún tipo de transformación en las imágenes mencionada, al efectuar la comparación, las correspondencia entre puntos característicos debe mostrar un cierto paralelismo y la distancia entre puntos debe ser más o menos equivalente dentro de unos ciertos márgenes. En este Proyecto se demuestra la gran mejora en los resultados obtenida al incluir este *post-procesado* en la implementación original de Lowe [8].

La etapa de comparación de los descriptores pone de manifiesto la elaboración de las primeras conclusiones de este proyecto. Estas se refieren a la

superioridad de los descriptores locales SIFT y SURF junto con el descriptor histograma HSV respecto del resto. La alternancia de las primeras posiciones en cuanto a rendimiento se refiere se produce entre los descriptores locales y el histograma HSV dependiendo de la categoría analizada; así en el caso de los cambios de ángulo o movimiento de objetos en la escena, es el histograma HSV quien domina la tabla de rendimientos; por el contrario, los descriptores locales SIFT y SURF relevan al anterior en el resto de casos, incluyendo el caso general en el que se representa la aparición de todas las transformaciones analizadas en este trabajo.

Finalmente en la última etapa de combinación, motivada por la esperanza de superar los resultados obtenidos de la etapa anterior, se evalúa el rendimiento de las distintas combinaciones de descriptores. Junto con esta etapa se presenta un estudio del coste computacional asociado a los descriptores en las diferentes tareas de creación y comparación de los mismos así como del espacio requerido para ser almacenados. Las conclusiones nacidas de esta última etapa cumplen con las expectativas y exponen a la combinación de los descriptores HSV-SURF como la mejor de las combinaciones, alcanzando unos resultados de más del 5 % sobre el mejor de los descriptores individuales, SIFT, y reduciendo el coste computacional asociado al mismo.

Además de ésta, se identifican otras mejoras como la combinación de otros descriptores que superan con altos márgenes el mejor de los rendimientos individuales de ambas componentes como es el caso de la combinación CL-RGB.

A través del trabajo aquí presentado se ha demostrado que la combinación de descriptores supone una mejora respecto de los resultados individuales en todas las categorías analizadas. Estos resultados pueden ser de gran utilidad para el desarrollo de aplicaciones en tiempo real en las que sea necesario la eliminación de redundancia o la detección de duplicados, como puede ser el caso de los resúmenes de video o la recuperación de contenido.

Queda como mención al posible trabajo futuro en esta línea la evaluación de combinaciones con un mayor número de descriptores así como con distintas ponderaciones de las componentes. La elección de otros refinamientos sobre los descriptores ya existentes como es el caso de PCA-SIFT o GLOH [34] que si bien proponen ciertas ventajas también suponen ciertos inconvenientes que sería recomendable incluir en el estudio comparativo.

Por último apuntar también en esta línea de trabajo futuro la creación de una base de datos más heterogénea en cuanto a la variedad dentro de cada categoría y también el incremento en el número de transformaciones analizadas, es decir, otros tipos de escenas como oclusiones, borrosidad en las imágenes o diferentes compresiones de los vídeos, para valorar así los distintos descriptores. El estudio comparativo ha de contar por otro lado con un mayor número de descriptores de manera que sean analizados simultáneamente y con más objetividad los descriptores elegidos.

Glosario de acrónimos

- **CBIR:** Content-Based Image Retrieval
- **CLD:** Color Layout Descriptor
- **CRLG:** Correlograma
- **DCT:** Discrete Cosine Transform
- **DoG:** Diference of Gaussians
- **GLOH:** Gradient Location and Orientation Histogram
- **HSV:** Hue Saturation Value
- **MPEG:** Moving Picture Experts Group
- **PR:** Precision-Recall
- **RGB:** Red Green Blue
- **ROC:** Receiver Operator Characteristic
- **SIFT:** Scalable Invariant Feature Transform
- **SURF:** Speeded Up Robust Features
- **TRECVID:** Video Retrieval Evaluation TREC

Herramientas utilizadas

El presente documento ha sido elaborado por el autor utilizando LaTeX. El formato de texto es Computer Roman Modern de tamaño 11pt. Todos los gráficos e imágenes han sido incluidos en formato Encapsulated Post Script

Notas sobre el copyright®

Los derechos de cualquier marca registrada mencionada en el presente documento son propiedad de sus respectivos autores.

Referencias

[1] F. Viksten, P. Forssen, B. Johansson, and A. Moe, "Comparison of local image descriptors for full 6 degree-of-freedom pose estimation," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pp. 2779–2786, IEEE, 2009.

[2] J. Feng, "Combining minutiae descriptors for fingerprint matching," *Pattern Recognition*, vol. 41, no. 1, pp. 342–352, 2008.

[3] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, "Video copy detection: a comparative study," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, pp. 371–378, ACM, 2007.

[4] T. Deselaers, D. Keysers, and H. Ney, "Features for image retrieval: An experimental comparison," *Information Retrieval*, vol. 11, no. 2, pp. 77–107, 2008.

[5] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, pp. 1349 –1380, Dec. 2000.

[6] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, pp. 177–280, July 2008.

[7] M. Swain and D. Ballard, "Color indexing," *International journal of computer vision*, vol. 7, no. 1, pp. 11–32, 1991.

[8] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] "Text of iso/iec 15938-3 multimedia content description interface - part 3: Visual." Final Committee Draft. Document No. N4062. Singapore. March 2001.

[10] K. Mikolajczyk, "Detection of local features invariant to affine transformations," *PhD hesis, Institut National Polytechnique de Grenoble, France*, 2002.

[11] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," 2004.

[12] T. Deselaers, D. Keysers, and H. Ney, "Features for image retrieval: A quantitative comparison," *Pattern Recognition*, pp. 228–236, 2004.

[13] J. Puzicha, J. Buhmann, Y. Rubner, and C. Tomasi, "Empirical evaluation of dissimilarity measures for color and texture," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2, pp. 1165–1172, IEEE, 2002.

[14] T. Randen and J. Husoy, "Filtering for texture classification: A comparative study," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 4, pp. 291–310, 2002.

[15] G. Carneiro and A. Jepson, "Multi-scale phase-based local features," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1, IEEE, 2003.

[16] P. Moreels and P. Perona, "Evaluation of features detectors and descriptors based on 3d objects," *International Journal of Computer Vision*, vol. 73, no. 3, pp. 263–284, 2007.

[17] K. Mikolajczyk, B. Leibe, and B. Schiele, "Local features for object class recognition," 2005.

[18] A. Majumdar and R. Ward, "Discriminative sift features for face recognition," in *Electrical and Computer Engineering, 2009. CCECE '09. Canadian Conference on*, pp. 27 –30, May 2009.

[19] L. Kotoulas and I. Andreadis, "Colour histogram content-based image retrieval and hardware implementation," *Circuits, Devices and Systems, IEE Proceedings* -, vol. 150, no. 5, pp. 387–93, 2003.

[20] S. Siggelkow, *Feature histograms for content-based image retrieval*. PhD thesis, Albert-Ludwigs-Universität Freiburg, 2002.

[21] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, (New York, NY, USA), pp. 357–360, ACM, 2007.

[22] A. Murillo, J. Guerrero, and C. Sagues, "Surf features for efficient robot localization with omnidirectional images," in *Robotics and Automation, 2007 IEEE International Conference on*, pp. 3901 –3907, 2007.

[23] C. Valgren and A. Lilienthal, "Sift, surf and seasons: Long-term outdoor localization using local features," in *Proceedings of the European Conference on Mobile Robots (ECMR 07)*, Citeseer, 2007.

[24] J. Han and K. Ma, "Fuzzy color histogram and its use in color image retrieval," *Image Processing, IEEE Transactions on*, vol. 11, no. 8, pp. 944–952, 2002.

[25] J. Domke and Y. Aloimonos, "Deformation and viewpoint invariant color histograms," in *British Machine Vision Conference*, vol. 2, pp. 509–518, 2006.

[26] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pp. 762–768, IEEE Computer Society, 1997.

[27] L. Cieplinski, "Mpeg-7 color descriptors and their applications," in *Computer Analysis of Images and Patterns*, pp. 11–20, Springer, 2001.

[28] M. Stricker and A. Dimai, "Color indexing with weak spatial constraints," *Storage and Retrieval for Image and Video Databases IV*, vol. 2670, 1996.

[29] G. Pass and R. Zabih, "Histogram refinement for content-based image retrieval," in *Proceedings 3rd IEEE Workshop on Applications of Computer Vision, 1996. WACV'96.*, pp. 96–102, 1996.

[30] P. Rousseeuw and A. Leroy, *Robust regression and outlier detection*. John Wiley & Sons Inc, 1987.

[31] J. Hafner, H. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 7, pp. 729–736, 2002.

[32] D. Haussler, "Decision theoretic generalizations of the pac model for neural net and other learning applications," *Information and computation*, vol. 100, no. 1, pp. 78–150, 1992.

[33] M. Brown and D. Lowe, "Invariant features from interest point groups," in *British Machine Vision Conference, Cardiff, Wales*, pp. 656–665, Citeseer, 2002.

[34] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *Proceedings of the ninth European Conference on Computer Vision*, May 2006.

[35] T. Lindeberg, "Scale-space for discrete signals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 234–254, 1990.

[36] P. Simard, L. Bottou, P. Haffner, and Y. Le Cun, "Boxlets: a fast convolution algorithm for signal processing and neural networks," *Advances in Neural Information Processing Systems*, pp. 571–577, 1999.

[37] K. Derpanis, "Integral image-based representations," *Department of Computer Science and Engineering, York University, Paper*, 2007.

[38] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2002.

[39] F. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *Proceedings of the Fifteenth International Conference on Machine Learning*, vol. 445, 1998.

[40] C. Drummond and R. Holte, "Explicitly representing expected cost: An alternative to roc representation," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 198–207, 2000.

[41] C. Drummond and R. Holte, "What roc curves can not do (and cost curves can)," in *Proceedings of the ROC Analysis in Artificial Intelligence, 1st International Workshop*, pp. 19–26, 2004.

[42] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," *Advances in Information Retrieval*, pp. 345–359, 2005.

[43] L. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, ACM, 2006.

[44] M. Goadrich, L. Oliphant, and J. Shavlik, "Learning ensembles of first-order clauses for recall-precision curves: A case study in biomedical information extraction," *Inductive Logic Programming*, pp. 421–456, 2004.

[45] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: overview and proposals," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 593 – 601, 2001.

[46] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

[47] S. Jeong, "Histogram-based color image retrieval," *Psych221/EE362 Project Report, Stanford University*, 2001.

[48] U. Park, S. Pankanti, and A. Jain, "Fingerprint verification using sift features," in *Proceedings of SPIE Defense and Security Symposium*, 2008.

[49] J. J. Foo and R. Sinha, "Pruning sift for scalable near-duplicate image matching," in *Proceedings of the eighteenth conference on Australasian database - Volume 63*, ADC '07, (Darlinghurst, Australia, Australia), pp. 63–71, Australian Computer Society, Inc., 2007.

ANEXO I Métricas L_1 y L_2

Las normas o distancias L_1 y L_2 son una particularización de la norma L_p, siendo $p > 1$ y $p \in \mathbb{N}$.

La distancia L_p en un espacio de dimensión n entre dos puntos $s = (s_1, s_2, \ldots, s_n)$ y $t = (t_1, t_2, \ldots, t_n)$ se define como:

$$d(s,t) = \|s - t\|_p = \left(\sum_{i=1}^{n} \mid s_i - t_i \mid^p \right)^{\frac{1}{p}}$$

En el caso particular de $p = 1$ y $p = 2$ se obtienen las distancias L_1 y L_2 respectivamente.

La **norma** L_1 también conocida como norma Taxicab, distancia Manhattan, distancia Cityblock o simplemente distancia L_1, se calcula como la suma de las diferencias absolutas de las coordenadas de los puntos. Respecto de la definición anterior se obtiene:

$$\|s - t\| = \sum_{i=1}^{n} \mid s_i - t_i \mid$$

El nombre de Taxicab o Manhattan está relacionado con la distancia que un taxi ha de recorrer en un grid de calles rectangular para ir del punto s al punto t. Este grid tiene en la morfología de la isla de Manhattan a uno de sus máximos exponentes.

El conjunto de vectores cuya norma L_1 es una constante dada forman la superficie de un politopo de dimensión igual a la de la norma menos 1.

En el caso de la **norma** L_2, conocida como la distancia Euclidea, se calcula como:

$$\|s - t\|_2 = \sqrt{\sum_{i=1}^{n} (s_i - t_i)^2}$$

En el espacio $\mathbb{R}^2$ o espacio Euclídeo, la distancia Euclidea se refiere a la distancia comúnmente conocida entre dos puntos en el plano $x \perp y$ representada por el segmento que los une.

El conjunto de vectores cuya norma Euclidea es una constante dada forman la superficie de una n-esfera, siendo n la dimensión del espacio Euclídeo.

ANEXO II Resultados de las comparaciones

A continuación se muestran los resultados obtenidos sobre la relación de imágenes que algunos de los descriptores presentan en función de la imagen original o query correspondiente en un escenario global.

Concretamente se presentan los resultados generados por el descriptor SIFT, cuyo rendimiento supera al resto de descriptores de forma individual, y los resultados de los descriptores histograma HSV y SURF tanto de forma individual como colectiva, siendo esta combinación la mejor de todas las producidas.

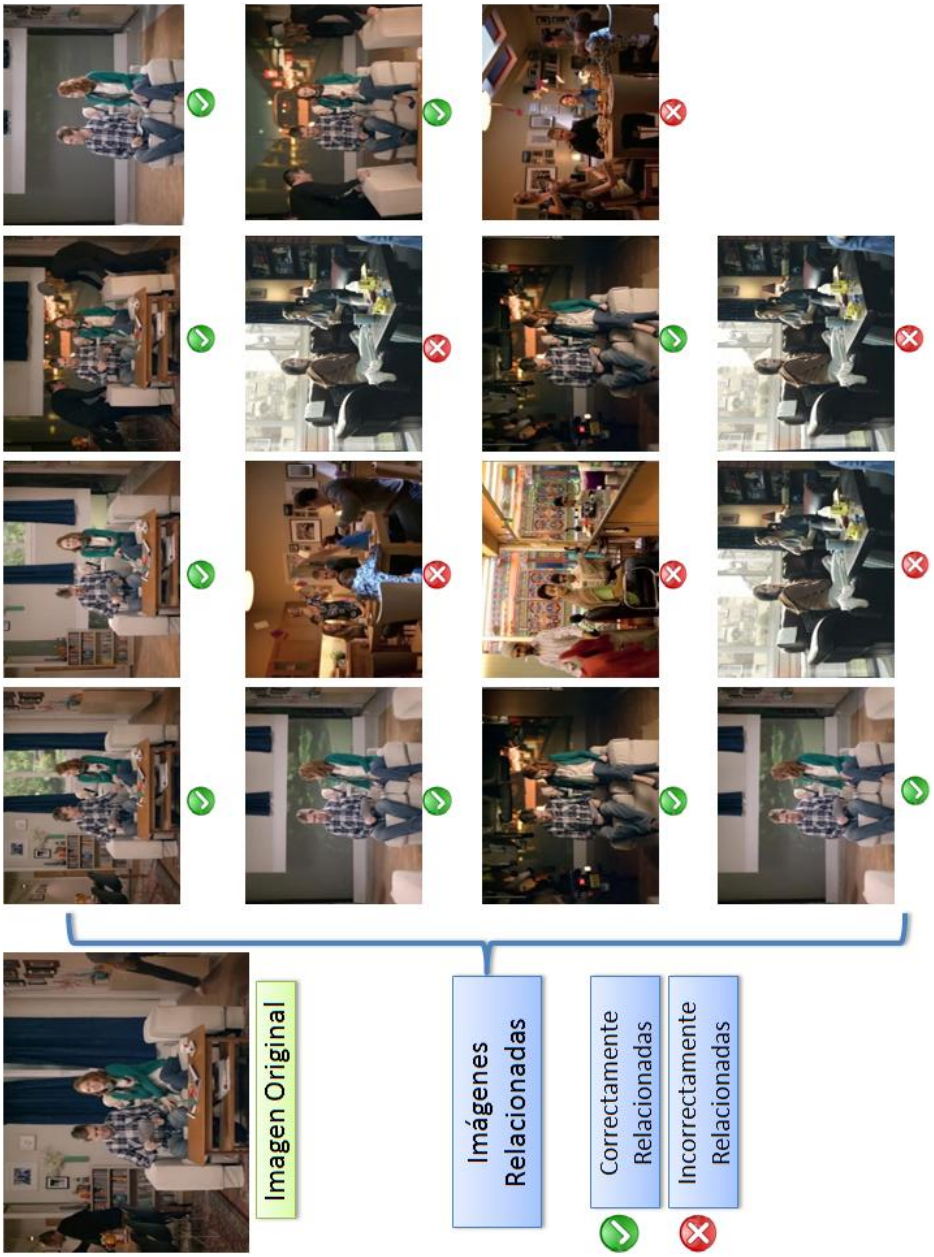

Figura 5.1: Ejemplo de imágenes relacionadas descriptor SIFT

Figura 5.2: Ejemplo de imágenes relacionadas descriptor HSV

Figura 5.3: Ejemplo de imágenes relacionadas descriptor SURF

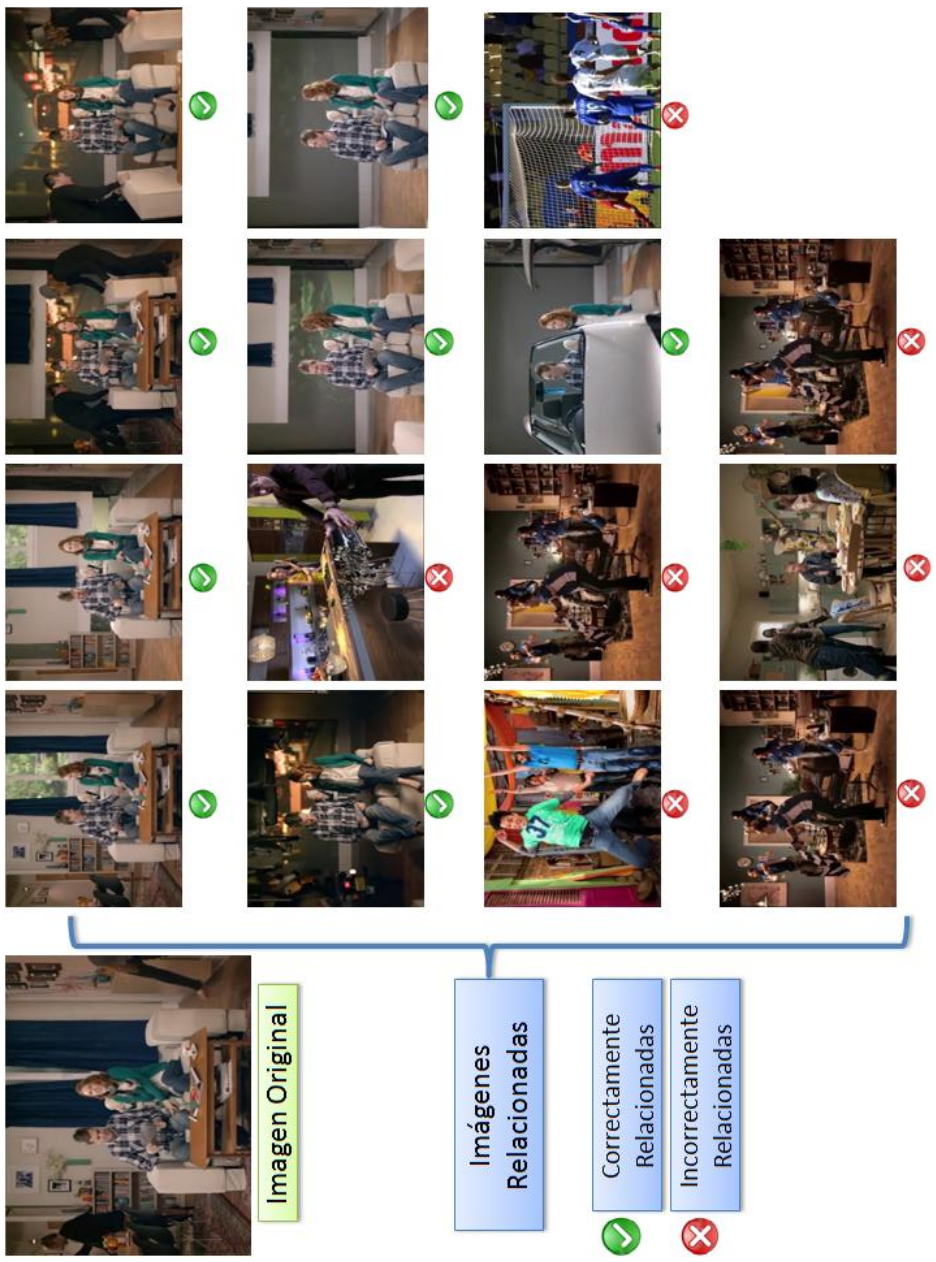

Figura 5.4: Ejemplo de imágenes relacionadas combinación descriptores HSV-SURF

Presupuesto

1) Ejecución Material

- Compra de ordenador personal (software incluido)..... 2.000 €
- Alquiler de impresora láser durante 6 meses ..... 210 €
- Material de oficina ..... 45 €
- Total de ejecución de material ..... 2.255 €

2) Gastos Generales

- 18 % sobre Ejecución Material ..... 405,9 €

3) Beneficio Industrial

- 6 % sobre Ejecución Material ..... 135,3 €

4) Honorarios Proyecto

- 800 horas a 15 € / hora ..... 12.000 €

5) Material Fungible

- Gastos de impresión ..... 180 €
- Encuadernación .....135 €

6) Subtotal Presupuesto

- Subtotal del presupuesto ..... 15.111,2 €

7) I.V.A. aplicable

- 18 % Subtotal Presupuesto ..... 2.720,02 €

8) Total Presupuesto

- Total Presupuesto ..... 17.831,22 €

Madrid, Febrero 2011

El Ingeniero Jefe de Proyecto

Fdo.: Óscar Boullosa García

Ingeniero Superior de Telecomunicación

Pliego de condiciones

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de una *Estudio comparativo de descriptores visuales para la detección de escenas cuasi-duplicadas*. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales.

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contrata" anteriormente llamado "Presupuesto de Ejecución Material"que hoy designa otro concepto.

Condiciones particulares.

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.

Publicaciones

A Comparative Study of Visual Descriptors for Near-Duplicate Scene Detection

Oscar Boullosa, Víctor Valdés, José M. Martínez
Video Processing and Understanding Lab, Universidad Autónoma de Madrid
oscar.boullosa@estudiante.uam.es, {victor.valdes, josem.martinez}@uam.es

Abstract

In this paper we present a comparative study of visual descriptors for near-duplicate scene detection within video content. The main goal is to measure the performance of several widely used image descriptors according to different criteria, namely their capacity to identify similar scenes when changes in the point of view, illumination, object movement and camera zoom occur. Afterwards, a study combining different descriptors is carried out and a combination of both computationally efficient and discriminative descriptors are proposed. The outcomes of this study can be applied for video summarization or retrieval approaches based on visual descriptors in order to select the most appropriate descriptors.

1. Introduction

In the last decade there has been a huge increase in multimedia information due to the massive use of the video technology in real life. As a result there is a growing need for video summarization and retrieval techniques in order to manage these contents and to make them available in a more efficient way. We may mention as examples large movie databases such as Youtube, movie sellers who may wish to automatically create movie trailers or news agencies or broadcasting companies with huge video collections which could create brief summaries or search for content in a more efficient way.

The mentioned techniques heavily rely on visual comparisons aimed for the elimination of redundant information or retrieve similar video segments. Such visual comparisons are carried out making use of many different existing visual descriptors. However, such visual descriptors have been usually evaluated focusing on their performance for the identification or retrieval of similar content considering a semantic point of view and not focusing on their performance for detecting near-duplicate scenes or retakes. This paper focuses in the evaluation of visual descriptors in those situations: the detection of similar scenes under different viewing conditions such as illumination and viewpoint changes, zoom variations and object or background movements.

Different works have evaluated visual descriptors in the context of matching and recognition. Carneiro and Jepson [3] evaluated the performance of local descriptors using Receiver Operating Characteristics (ROC). Randen and Husoy [10] compared different descriptors for one texture classification algorithm, and image pairs were used to compare the descriptors. Mikolajczyk and Schmid [8] exhaustive evaluation used the precision-recall criterion to examine descriptors under several image transformations.

These works have been carried out comparing the performance of local descriptors individually, however no comprehensive study has been carried out to analyze the performance of these techniques when dealing with different viewing conditions in the context of near-duplicate scenes. Moreover, in this work, the performance of different descriptor combinations is carried out, aiming to determine their performance in comparison with single descriptors.

In this paper, we compare the performance of four global descriptors: (1) RGB Histogram [7]; (2) HSV Histogram [7]; (3) Color Layout [7, 1]; and (4) Color Correlogram [5]; and two local descriptors: (5) Scale-Invariant Feature Transform (SIFT) [6]; and (6) Speeded-Up Robust Features (SURF) [2]. These descriptors, covering an interesting range of different approaches, have been selected due to their widely spread usage and popularity and because they cover a wide range of computational costs . First of all, we carry out an individual optimization of each descriptor, finding the calculation parameters providing best results. Then, with the optimized set of parameters, a comparative study among the different descriptors is performed. Finally, an exhaustive test combining the descriptors in pairs is carried out to determine if an improvement of the results is possible. The evaluation criterion is the area under the precision-recall curve detailed in [4][9], i.e., the number of correct and false matches between two images.

The remainder of this paper is organized as follows: In Section 2 we briefly overview the selected visual descriptors, Section 3 describes the content set used for the evalu-

ation. Experimental results are presented in Section 4 and, finally, the work is concluded and future work is proposed in Section 5.

2. Visual Descriptors

In this section an overview of the evaluated descriptors is presented. We classify the selected descriptors in two groups, global and local descriptors, attending to their application over the whole image or over a set of keypoints computed by a detector (in this work SIFT [6] and SURF [2] detectors are applied).

2.1. Global Descriptors

- *Color Histogram* [7]*:* The color histogram contains the frequency of occurrence of each color in the image stored in the bins of the histogram. The more bins a color histogram contains the more discrimination power it has, however a large number of bins will not only increase the computational cost of image comparisons, but will also reduce the comparisons efficiency (e.g. too many bins make histograms to be more sensible to illumination changes). For comparing color histograms we used 5 different distance metrics that are specified in section 4.

- *Color Layout:* The color layout descriptor -CLD- [1, 7] is a compact descriptor that uses representative colors on an 8x8 grid followed by a Discrete Cosine Transform (DCT). The color space adopted for CLD is YCrCb. The distance metric used for to perform the comparisons of color layout descriptors between two images $I \; and \; I'$, is given by:

$$\mid I - I' \mid = 0.2\,(diff_{\mathrm{Y}}) + 0.4\,(diff_{Cb}) + 0.4\,(diff_{Cr})$$

where *diff*$_Y$*, diff*$_{Cb}$ *and diff*$_{Cr}$ are the difference between the coefficients of the corresponding components.

- *Color Correlogram:* The color correlogram descriptor [5] can be presented as an extension of the color histogram that includes information about color spatial correlation. Intuitively, the color correlogram represents how the spatial correlation among colors in the image varies with respect to their distance *d*. The color correlogram is characterized as follows: Let *I* be an *N x M* image quantized in *m* colors bins $c_1, ..., c_m$, for a pixel $p = (x, y) \in I$, $p \in I_c$ means that the color of *p* is *c*. Then the correlogram of *I* is defined by:

$$\gamma^{(k)}_{c_i,c_j}(I) \triangleq \Pr_{p_1 \in I_{c_i}, p_2 \in I} \left[p_2 \in I_{c_j} \mid\mid p_1 - p_2 \mid = k \right]$$

where $(i, j) \in \{1, 2, ..., m\}$, $k \in \{1, 2, ..., d\}$ and $\mid p_1 - p_2 \mid$ is the distance between the pixels $p_1$and p_2.

The distance metric [5] used to compare the image correlograms $I \; and \; I'$ is given by:

$$\mid I - I' \mid_{\gamma} \triangleq \sum_{i,j \in [m], k \in [d]} \frac{\mid \gamma^{(k)}_{c_i,c_j}(I) - \gamma^{(k)}_{c_i,c_j}(I') \mid}{1 + \gamma^{(k)}_{c_i,c_j}(I) + \gamma^{(k)}_{c_i,c_j}(I')}$$

2.2. Local Descriptors

- *SIFT:* The SIFT descriptor developed by Lowe [6], describes each detected region as a unique feature by making use of the gradient information of the neighboring pixel intensities. The SIFT descriptor is a four step process: (1) scale-space extrema detection; (2) keypoint localization; (3) orientation assignment; and (4) keypoint descriptor. The first step identifies a set of potential interest points, detected by the region detectors. For each region, a model is fitted to determine its location and scale, the selection of the keypoints is then based on the measures of their stability. The orientation of the keypoints are then assigned, and lastly, descriptors are fitted using the image gradients.

- *SURF:* The SURF descriptor [2] proposed a number of improvements over SIFT, in particular, it focused on improving the repeatability, distinctiveness and robustness and, at the same time, decreased the computation required time. The descriptor is based on the distribution of the gradients, and simplified to the essential. Experimental work has shown that the SURF descriptor has a significant advantage in terms of performance and computation time when compared with the SIFT descriptor [2].

For *SIFT* and *SURF* descriptors we have proposed a common distance metric constituted by two weighted components: $d_{spatial}$, consisting in a spatial distance measure between matched points and $d_{matching}$, which corresponds to one of the three applicable matching strategies described in the next paragraph. The metric formula is as follows:

$$distance = [(w_1 * (1 - d_{matching})) + (w_2 * d_{spatial})]$$

where $(w_1, w_2) \in [0, 1]$ are the weights for each distance component.

Three modalities for the matching distance, $d_{matching}$, have been evaluated. In the case of *threshold-based matching*, two keypoints are matched if the euclidean distance between their descriptors is below a $threshold_1$ and, in addition, each descriptor can only have one match. In the case of *nearest neighbor-based matching*, the descriptors of the two

nearest neighbor keypoints B and C for the keypoint A are selected. If the distance ratio between the descriptors D_B and D_C respectively to the descriptor D_A is below a threshold, then the keypoints A and B are matched, that means, if $\| D_B - D_A \| / \| D_C - D_A \| < threshold_2$. The third matching strategy, *unique minimum-distance matching*, consists of assigning the minimum distance descriptor to each unique match. All matching strategies compare each keypoint of the reference image with each keypoint of the transformed image.

3. Performance Evaluation

The descriptor comparison study was carried out making use of real images from a video collection created for such purpose. The data set was selected aiming to include examples of different kind of scene variations and has been created from different types of video such as BBC rushes videos (MPEG-1) from the TREC Video Retrieval Evaluation content set[1] or Time-Slice® Films's videos from Vimeo[2] (a detailed description can be found at the following web site http://www-vpu.eps.uam.es/publications/XXXX). The image size is 352 x 288 pixels as it is the minimum and most representative size of the selected videos, resizing the rest. Several interesting frames from each video were manually selected from each original video in order to obtain the test data set which can be found in the above mentioned web page.

The data set consists of 4 categories (viewpoint change, illumination variation, position change and zoom change) with 500 images each. Within each category there are different interrelated image groups, each one composed by a variable number, between 5 and 25, of images representing the same scene but affected by some transformation. View point change sequences represent a camera position movement from a front-parallel view to one with significant foreshortening at approximately 45-50 degrees in both directions. The illumination changes are produced by varying the main light source position or different illumination conditions. The position change scenes describes different trajectories of the objects on the foreground and background in the pictures and finally zoom pictures are obtained by varying the optical distance between the camera and the objects in the scene. An example of each of the referred categories is shown in Fig.1.

To accomplish the comparative descriptor evaluation we used the *area under precision-recall* criterion (AUPRV) described in [4][9] as a metric to evaluate the matching algorithm performances, which is based on the number of correct and false matches obtained for any image comparison. As we have mentioned, the data set is formed by interrelated images groups. Consequently when a comparison between two images is performed, the result is considered as a correct match if the reference and compared images belong to the same group or a false match otherwise.

[1] http://trecvid.nist.gov/

[2] http://vimeo.com/timeslice

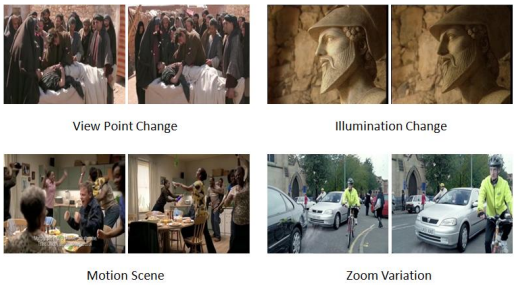

Figure 1. Scene Types

4. Experimental Results

We designed three sets of experiments (Experiment sets A, B and C) for the evaluation of the descriptors performance. Experiment set A focuses on the visual descriptors individually, determining which of the metrics and parameters proposed for each descriptor performs better on average for the different kinds of near-duplicate scenes. In Experiment set B, a comparative study among all the optimized descriptors is performed in order to find out the best descriptor for each situation and a global average performance for all types of near-duplicate scenes. Experiment set C is carried out by combining all the descriptors in pairs, trying to improve the results obtained in the Experiment set B.

Finally, a discussion about the computational cost of the descriptors is presented in order to evaluate the tradeoff between the descriptors accuracy and cost.

4.1. Optimization

An optimization stage with different parameters and matching metrics has been made for the tested descriptors, except the Color Layout which is defined as a MPEG-7 standard. We used 25 different reference images which were compared with 150 random transformed images each for all descriptors in this stage. The complete results can be found in the previous mentioned web page as we present in this paper only the optimal results for each descriptor.

Regarding to *Color histograms (RGB and HSV)*, we have analyzed five different metrics: *Bhattacharyya, Chi-square, Cityblock, Euclidean and Intersection*. All of them were applied making use of both whole image histogram and over a combination of four histograms per image, obtained from the four regular rectangular regions of the images. The aim

was to analyze the best performance for each particular image transformation and to obtain an optimal configuration for a general situation (by computing the average results for the four types of image transformations). The best results are listed in Table 1.

For *Color Correlogram*, we have carried out different resizes of the image before creating the correlogram. The considered options were: *no resize, 200 x 200, 100 x 100*, and *50 x 50* pixels, being the size of the images 352 x 288 pixels. The best results are shown in Table 1.

The optimization of the *SIFT* and *SURF* descriptors was carried out by testing all the possible w_1 and w_2 values within the [0.1, 0.9] range with a step of 0.1, $threshold_1$ within the possible values $(0.33, 0.43, 0.5, 0.6, 0.66, 0.73)$ and $threshold_2$ within the $(0.60, 0.65, 0.70, 0.75, 0.80)$ values. The best results are shown in Table 1.

From the depicted results we may observe that, for both histograms, the *CityBlock over four regular rectangular regions* distance is optimal not only for a general scenario including all the types of scene variations but also for each of them individually. At the same time the HSV color space obtains better results than the RGB one in all situations.

For the *Color Correlogram* non-resizing the original image before its calculation is the best option; compacting the image makes it lose discriminative information and therefore lower descriptor performance is achieved.

For SIFT and SURF descriptors similar results are obtained. In both cases the optimal distance is achieved with the threshold-based matching strategy and $w_1 = w_2 = 0.5$ but with a $threshold_1 = 0.6$ value for SIFT and a $threshold_1 = 0.33$ value for the SURF descriptor. Therefore, both descriptors share the same weighed terms composition and the same matching strategy but the SIFT descriptor seems to be more restrictive due to the higher threshold value.

4.2. Descriptor Comparison

In this section, we evaluate the descriptor performance for different types of scenes making use of the the selected configurations described above. The results were calculated by evaluating each descriptor and each scene type making use of 100 different reference images which were compared with 200 random transformed images each. The results are depicted in Table 2.

For motion scene and view point change transformations the *HSV histogram* obtains the best results with a slightly better performance over the rest of descriptors. Regarding to illumination changes there is a great performance difference between *SIFT* and *SURF* and the rest of descriptors. This is caused because this transformation affects to the overall pixel values and the color descriptors are usually more sensible to this kind of changes. On the other hand, the keypoint based descriptors, SIFT and SURF, are specifically designed to deal with this kind of transformations. The same reasoning can be applied to zoom change scenes but, in this case, the *HSV histogram* works almost as good as the *SIFT* and *SURF* descriptors. We may also realize that all descriptors have a similar curve shape but a smaller overall robustness can be observed for motion scenes. Finally we can check that, for a general situation (combining all the four types of transformations), S*IFT, SURF* and *HSV histogram* (with a slightly worse performance) descriptors, perform significantly better than the rest of the descriptors.

4.3. Combined Descriptors Comparison

The aim of this section is to discuss the results obtained for all the pairwise combinations of descriptors pursuing to improve previously obtained results for individual descriptors. As we are performing the study over 6 descriptors, a total of 15 different combinations of local and global descriptors were carried out. For each pair of descriptors the combined descriptor distance is calculated as a weighted mixture of each descriptor distance as $CombineDistante = [(w * dist_1) + ((1 - w) * dist_2)]$ where $w \in [0, 1]$ is the weight and $dist_i \in [0, 1]$ are the normalized distances between reference and compared images of each combined descriptor.

The experiment is performed by all the possible w values within$[0.1 : 0.9]$ with a step of 0.1. Due to the impossibility of showing all the tables for each w tuning value we present the results for the best selection of the weighting parameter which corresponds to $w = 0.2$ and also only the best 5 descriptor combinations are depicted in Table 3. All the complete Tables can be found in the previous mentioned web site.

As we may see in table 3, the best results for every type of scene are obtained with the HSV Histogram combined with the SIFT or SURF descriptors. These results rely on the ability of SIFT and SURF descriptors to identify similar scenes under different transformations with the HSV Histogram contribution in the cases when they are vulnerable to view point changes. We may observe that, although the contribution of HSV Histogram descriptor is small (with a weight of only 0.2) the combined descriptor performs very well in all the possible types of transformations, including the general case (evaluation with all the types of transformations). In addition the results of the combined descriptor improve the performance obtained from the individual descriptors in almost a 5%.

4.4. Computational Cost

The computational cost of the different descriptors is another aspect to be taken into consideration when the per-

SceneType - Ranking		RGB Histogram	HSV Histogram	Color Correlogram	SIFT	SURF
View Point Change	Distance	$Bhattacharyya_1$	$Bhattacharyya_1$	$Correlogram_1$	$C_1 \mid 0.7 \mid 0.66$	$C_1 \mid 0.5 \mid 0.33$
	AUPRV	0.9663	0.9801	0.9597	0.9874	0.9925
Illumination Change	Distance	$Cityblock_2$	$Cityblockb_2$	$Correlogram_{50}$	$C_1 \mid 0.5 \mid 0.6$	$C_1 \mid 0.5 \mid 0.33$
	AUPRV	0.7718	0.8343	0.6939	0.9945	0.9755
Motion Scene	Distance	$Cityblock_2$	$Intersec_2$	$Correlogram_1$	$C_1 \mid 0.5 \mid 0.6$	$C_1 \mid 0.5 \mid 0.33$
	AUPRV	0.8444	0.8581	0.7989	0.7764	0.6994
Zoom	Distance	$Cityblock_2$	*Chi-square*$_2$	$Correlogram_{200}$	$C_1 \mid 0.5 \mid 0.43$	$C_2 \mid 0.5 \mid 0.8$
	AUPRV	0.8357	0.9055	0.7635	1	0.9951
General Scene	Distance	$Cityblock_2$	$Cityblock_2$	$Correlogram_1$	$C_1 \mid 0.5 \mid 0.6$	$C_1 \mid 0.5 \mid 0.33$
	AUPRV	0.8477	0.889	0.798	0.9385	0.9082

Table 1. Optimization Descriptor Results

The Area Under Precision-Recall Value (AUPRV) for the optimized descriptors. The subscripts 1 or 2 denotes whole or four regular rectangular regions of the images for RGB and HSV Histograms. In the case of Color Correlogram the subscript means the resize of the image, where 1 means no resize, and the rest values mean the resize applied. Finally, for SIFT and SURF descriptors C_i, with $i = (1, 2, 3)$, denotes the used $d_{matching}$ strategy (see section 2.2), second term refers to the distance component weight and third term is related to the threshold value.

SceneType - Ranking		1º	2º	3º	4º	5º	6º
View Point Change	Descriptor	HSV Histogram	Correlogram	RGB Histogram	Color Layout	SIFT	SURF
	AUPRV	0.9748	0.9428	0.9426	0.9376	0.9268	0.9
Illumination Change	Descriptor	SIFT	SURF	HSV Histogram	RGB Histogram	Color Layout	Correlogram
	AUPRV	0.9506	0.9258	0.7114	0.6389	0.5434	0.4326
Motion Scene	Descriptor	HSV Histogram	SIFT	SURF	RGB Histogram	Color Layout	Correlogram
	AUPRV	0.8905	0.847	0.8397	0.8294	0.8198	0.7848
Zoom	Descriptor	SURF	SIFT	HSV Histogram	Color Layout	RGB Histogram	Correlogram
	AUPRV	0.9682	0.9652	0.8816	0.8628	0.8378	0.7415
General Scene	Descriptor	SIFT	SURF	HSV Histogram	RGB Histogram	Color Layout	Correlogram
	AUPRV	0.9112	0.9084	0.8646	0.8122	0.7909	0.7254

Table 2. Descriptor Comparison

	Creation	Comparison	Descriptor size
HSV Histogram	130 ms.	6 ms.	384 bytes // 1536 bytes
HSV Histogram	290 ms.	5.4 ms.	96 bytes // 384 bytes
Color Layout	60 ms.	2.5 ms.	768 bytes
Correlogram	9800 ms.	5.5 ms.	960 bytes $\simeq$ 1 Kb
SIFT	7100 ms.	600 ms.	470 Kb
SURF	350 ms.	125 ms.	100 Kb

Table 4. Descriptor Computational Cost

formance of an image comparison technique is evaluated. The required computational effort for the calculation of the descriptor and the time required for comparing two images with that descriptor may imply the choice of a different descriptor for its implementation in real applications. The computational cost refers not only to the necessary time to perform the comparisons (and create the descriptors) but also to the required space for storing the descriptor as well as for carrying out the comparisons. Table 4 depicts the computational cost obtained for the selected image descriptors. In order to calculate the descriptor size, 4 bytes for each double type data is required to store the results because all the results are represented as doubles.

Attending to the time required for image comparison there are two different groups: global descriptors with a comparison time around (or less than) 1 hundredth of a second and local descriptors, that require several tenths of a second. Regarding the size of descriptors there are also two clearly differentiated groups: global descriptors require a low storage space, less than 1Kb, while local descriptors require much more storage, specially the SIFT descriptor.In the case of the SIFT and SURF descriptors, the results are averaged because the number of keypoints depends on both the dimensions and content of the images. In our case the size of the images is 288 x 352 pixels.

Taking all this into account, we can conclude that the computational cost of local descriptors might by up to 50 times above the required cost for global descriptors in the case of the SURF descriptor and above 200 times higher in the SIFT case. In fact and taking into account the results shown in the previous section, we may observe that the combination of HSV histogram and SURF descriptors achieves better results than the best single descriptor SIFT not only regarding the discriminative performance but also the computational cost efficiency.

5. Conclusions and future work

In this paper we have presented an experimental evaluation of several well-known global and local descriptors for matching and image detection of the same scene from a new content-based dataset consisting of retakes, view point changes, motion scenes, illumination and zoom variations.

An individual analysis for each type of scene was performed identifying which descriptor works better and afterwards a general situation analysis was carried out considering all the previous transformations over the images. In most of the analyzed situations, the SIFT descriptor obtains the best results, closely followed by SURF. This shows the

SceneType - Ranking		1º	2º	3º	4º	5º
View Point Change	D.Combination	HSV,SURF	HSV,SIFT	RGB,SURF	CL,HSV	RGB,SIFT
	AUPRV	0.9882	0.9873	0.976	0.9759	0.9754
Illumination Change	D.Combination	HSV,SIFT	CRLG,SIFT	HSV,SURF	RGB,SIFT	SIFT,SURF
	AUPRV	0.9609	0.9537	0.9514	0.9434	0.9409
Motion Scene	D.Combination	HSV,SURF	HSV,SIFT	RGB,SIFT	CRLG,HSV	RGB,SURF
	AUPRV	0.9056	0.9035	0.8925	0.891	0.8891
Zoom	D.Combination	HSV,SIFT	HSV,SURF	SIFT,SURF	RGB,SIFT	CRLG,SURF
	AUPRV	0.9946	0.9927	0.9738	0.9716	0.9702
General Scene	D.Combination	HSV,SURF	HSV,SIFT	RGB,SIFT	RGB,SURF	CRLG,SIFT
	AUPRV	0.9616	0.9595	0.9457	0.9409	0.9304

Table 3. Descriptor Combination Comparison

robustness and the distinctive character of the region-based descriptors. However, for view point change and motion scenes, the performance of the SIFT and SURF descriptors is below than HSV performance. Afterwards, given the high computational requirements of those local descriptors, specially for the SIFT case, they might not be the most appropriate approaches for all situations. In fact HSV histogram achieves quite well results, only a around 5% lower that the best single results belonging to SIFT descriptor, with much lower computational costs. Moreover, a further descriptor combination study has demonstrated that the combinations of HSV histogram with SURF and SIFT bring the best performances better in a 5% higher rate than the best individual performance of SIFT descriptor. Among the best combinational results we may conclude that the combination of HSV histogram with SURF descriptor is the most valuable combination as it performs similar to the HSV and SIFT combination while requiring a lower computational cost. In addition, other combinations without region-based descriptors perform very well and might be considered as an alternative when the high computational requirements of the local descriptors is an issue.

The presented results can be very useful for the development of real applications where the elimination of redundancy or detection of duplicates, for example in video summarization and retrieval, are required.

A more extensive and varied set of descriptors would complete the results shown in this work. Moreover, different descriptor combinations should be conducted in terms of more and different weighted components. It would be also interesting to accomplish this comparative study with scenes under other different viewing conditions such as occlusions, image blur or different video compression.

6. Ackowledgements

Work partially supported by the European Union under contract ICT-PSP-FP7-250527 (ASSETS): Advanced Search Services and Enhanced Technological Solutions for the Europeana Digital Library.

References

[1] Text of iso/iec 15938-3 multimedia content description interface - part 3: Visual. final committee draft, iso/iec/jtc/sc29/wg11, doc. n4062, march 2001.

[2] 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. Similarity Matching in Computer Vision and Multimedia.

[3] G. Carneiro and A. Jepson. Multi-scale phase-based local features. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–736 – I–743 vol.1, 2003.

[4] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[5] J. Huang, S. Ravi Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Spatial color indexing and applications. *International Journal of Computer Vision*, 35:245–268, 1999. 10.1023/A:1008108327226.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 10.1023/B:VISI.0000029664.99615.94.

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

[8] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615 –1630, 2005.

[9] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Performance evaluation in content-based image retrieval: overview and proposals. *Pattern Recognition Letters*, 22(5):593 – 601, 2001.

[10] T. Randen and J. Husoy. Filtering for texture classification: a comparative study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(4):291 –310, Apr. 1999.