Universitat Autònoma de Barcelona

MÁSTER DE INFORMÁTICA AVANZADA
MEMORIA DEL TRABAJO DE INVESTIGACIÓN
ITINERARIO CODIFICACIÓN, COMPRESIÓN Y SEGURIDAD

VISIÓN ESTÉREO. MODELO ANOVA DE BLOQUES ALEATORIZADOS

Autor: Jesús Jaime Moreno Escobar
Fecha: 9 de septiembre de 2008
Tutor: Rosa Maria Figueras i Ventura

Índice

Capítulo 1

Introducción

En los últimos años las imágenes tridimensionales se han vuelto parte de la vida cotidiana, popularizándose principalmente en películas de dibujos animados en tercera dimensión, aunque los grandes avances en este tipo de imágenes han sido en las de tipo volumétricas, las cuales dan apoyo a la visualización de ultrasonidos y resonancias magnéticas en la medicina [TPM03]. También se les puede encontrar en simuladores, los cuales preparan tanto a personal militar como civil en algunas situaciones dadas con un entorno conocido y por ende controlable. Este tipo de imágenes tridimensionales forman el eje z solamente proyectando la perspectiva de la visualización de una imagen bidimensional hacia un costado (figura 1.1), ayudándose de un punto de luz para crear la ilusión de sombra en contornos y con ello de distancias.

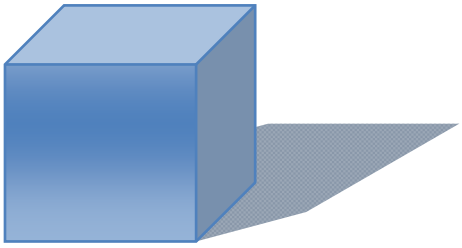


Figura 1.1.- Cubo.

Aunque el verdadero reto consiste en crear sistemas de visualización, los cuales obtengan los escenarios tridimensionales o estereoscópicos, como lo hace el sistema visual humano y así representar imágenes tridimensionales naturales de patrones no definidos. Con ello recibir eventos reales que puedan ser percibidos como si se estuviese en lugar donde se genera la transmisión. El reciente y creciente interés por las imágenes y los vídeos de escenas naturales en tres dimensiones, ha motivado a investigadores desarrollar novedosas técnicas de codificación [TPM03], tratamiento [VMP04], almacenamiento [SMT06] y transmisión [CKC04, MP04] de este tipo de imágenes. Cada una de dichas técnicas tiene ciertos detalles, por citar algunos, en cuanto a su procesamiento, es muy costoso procesar imágenes en tiempo real o en cuanto a su codificación, ya que se puede perder información de profundidad al ser comprimidas las imágenes. Como paso preliminar de cualquier sistema estereoscópico se requiere de un algoritmo que calcule la profundidad en la escena y es aquí donde se centrará la presente memoria.

Para poder plantear un algoritmo diferente hay que definir primeramente los objetos a alcanzar en este trabajo y estos son:

- Conocer el procedimiento de captura, representación computacional, codificación y visualización de imágenes naturales estereoscópicas.
- Exponer técnicas de estimación de profundidad.
- Proponer una técnica diferente a las actuales que estime la profundidad de una escena estereoscópica.

- Desarrollar un experimento que sustente la teoría del algoritmo propuesto.

La memoria consta de cuatro capítulos de los cuales el segundo y el tercero, son los que sustentan tanto teoría como práctica. En el segundo capítulo se abordará el proceso que siguen las imágenes naturales desde ser obtenidas hasta ser visualizadas por el espectador. La composición de este capítulo está pensada para llevar al lector por todo el proceso, empezando por el de la captura donde se le describirá como se calibran y ubican las cámaras y se obtiene un par de imágenes normalmente llamado par estéreo. Una vez que las imágenes han sido obtenidas se procederá a que observe como se calcula la profundidad en la escena, ya sea tomando regiones de la imagen izquierda y comparándolas con otras de la imagen derecha o también reuniendo características como colores, texturas, etc. de su par estéreo. Entonces la imagen resultante del cálculo de la profundidad o mapa de profundidad necesita ser codificada por ello se presentan tres técnicas de codificación de la tercera dimensión, la primera que se basa en encontrar patrones recurrentes multiescalares, mientras que la segunda y tercera se fundamentan en la Transformada Wavelet. Finalmente para apreciar la tercera dimensión se necesitarán de ciertos dispositivos, los cuales podrán ajustar automáticamente las imágenes tridimensionales a los ojos o requerirán de gafas de tipo polarizadas u obturadoras, por citar algunos tipos de estas.

En el tercer capítulo se revisarán cuatro técnicas que calculan las correspondencias que existen en una imagen derecha a partir de una izquierda y con ello estimar la profundidad. Esto podrá realizarse ya sea igualando bloques, correlacionado imágenes no rectificadas, basándose en encontrar la mejor igualación o programando de manera dinámica sin pasar por el mismo punto de la imagen dos veces. Visto y analizado todo lo anterior se propondrá un algoritmo alternativo, el cual ocupa un modelo estadístico que hace un análisis de varianzas (ANOVA) de bloques aleatorios de partes de la imagen izquierda y de la derecha. Con dichas partes o bloques, una vez comparados, se podrá saber que naturaleza tienen, es decir, si son exactamente iguales, que tan parecidos son, si son homogéneos o inclusive si son heterogéneos.

Para el análisis del estado del arte, la presente memoria se basó en metodología de [FCA07], la cual fue adaptada al tema. Primeramente se realizaron dos tareas, para profundizar la comprensión del diseño e implementación de sistemas estereoscópicos. La primera tarea comprendió en un análisis del diseño y de la implementación de imágenes en 3D, los modelos y los mecanismos de obtención de las mismas, donde las funciones que utilizan la geometría epipolar son las más frecuentemente utilizadas, mientras la segunda revisión fue similar pero se aplicó específicamente en estimar la profundidad de imágenes tridimensionales naturales.

Esta revisión comprendió alrededor de 35 artículos completos principalmente de los últimos 20 años, aunque existen algunos que datan de los años 1979, 1966 o inclusive del año 1838, los cuales contienen principalmente los temas del sistema visual humano, la geometría epipolar y algoritmos computacionales para la representación de imágenes estereoscópicas.

La búsqueda de información de los capítulos 2 y 3 de la presente memoria fue realizada en la base datos Xplore de la IEEE de Revistas y Congresos o donde se introdujeron las siguientes palabras clave en inglés:

1. 3d Image coding
2. Depth
3. Depth calculation
4. Stereo image coding
5. Stereoscopic image coding
6. Depth estimation
7. Depth Perception
8. Depth recognition
9. Epipolar Geometry
10. Stereoscopic depth

Mientras que para la generación de anexos se utilizó lo siguiente:

- La base de datos de la UPM, donde se buscó en los proyectos: Inspection, 3D y Visual Information Magement, se leyeron los resúmenes y se decidió cual descargar.
- Para temas generales y compresión de palabras también se buscó en:
  - Google
  - Scholar Google

- La Wikipedia.

El método de búsqueda fue hecho en varias etapas. Se comenzó leyendo los títulos y los resúmenes de la lista de artículos recuperados en las bases de datos anteriormente mencionadas. Se hizo un primer intento para resumir toda la información de dichos artículos en un diagrama holográfico (figura 1.2), donde se plasman los temas más importantes para pasar de lo más general como es la visión estereoscópica, discutiendo los temas que componen al sistema visual humano, la geometría epipolar y las representaciones o algoritmos computacionales hasta llegar a lo particular como una cámara, que es una analogía del sistema visual humano [MMA05, Mor05].

Además se realizó un modelo cibernético de un sistema de visión estereoscópica (figura 1.3). En dicho modelo se describe que para obtener una imagen en 3D se deberán realizar 6 etapas, las primeras corresponden a etapas con las cámaras; en las que se ubica el objeto, se calibran las cámaras y se capturan las imágenes bidimensionales. Después dichas imágenes pasan a un proceso de rectificación, cuando este finalice se realizan las correspondencias entre imágenes utilizando la geometría epipolar y se codifican según el algoritmo o técnica computacional necesitada. Con todo lo anterior se puede entonces reconstruir una escena tridimensional.

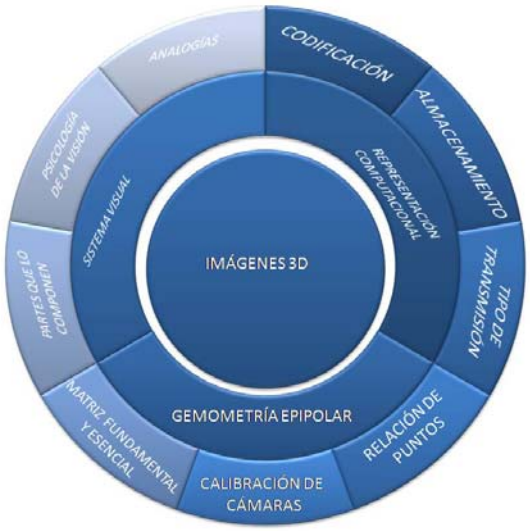


Figura 1.2.- Diagrama holográfico para imágenes tridimensionales.

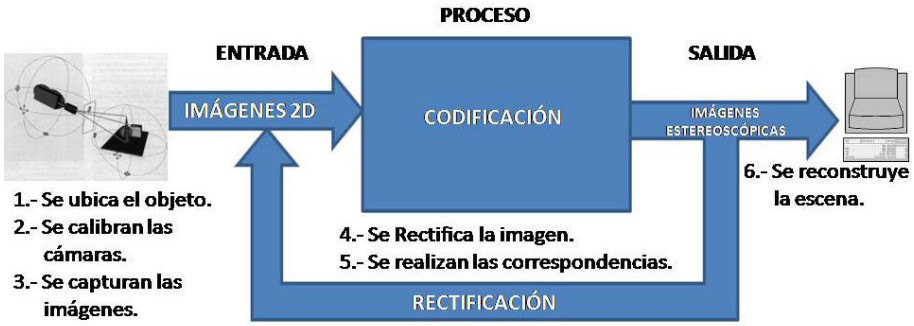


Figura 1.3.- Modelo cibernético de un sistema estereoscópico.

Se incluyeron artículos que describan al menos uno de los siguientes temas:

- Sistema visual humano y sus analogías mecánicas.
- Geometría epipolar, calculo y distribución de cámaras.
- Representaciones computacionales, las cuales pueden incluir algoritmos o aplicaciones desarrolladas.

Los artículos se buscaron específicamente en el tema del visión estereoscópica de imágenes naturales quedando excluidos en esta primera revisión aquellos que trataban a imágenes generadas por ordenador pero incluidos siguientes revisiones. De los artículos seleccionados se trató de registrar y leer los artículos completos.

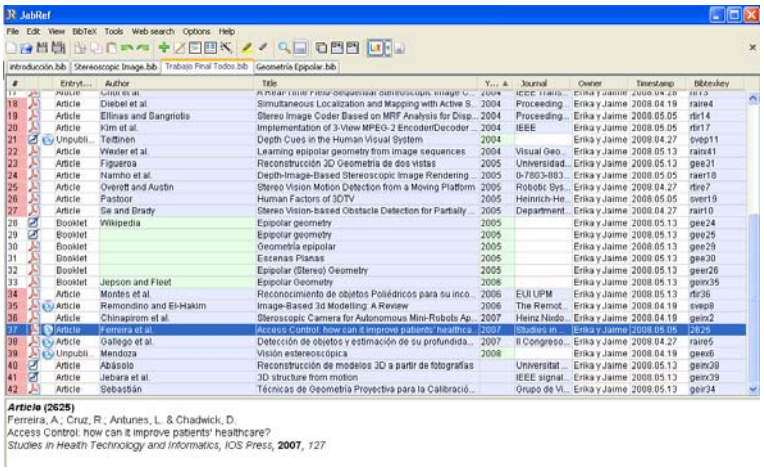
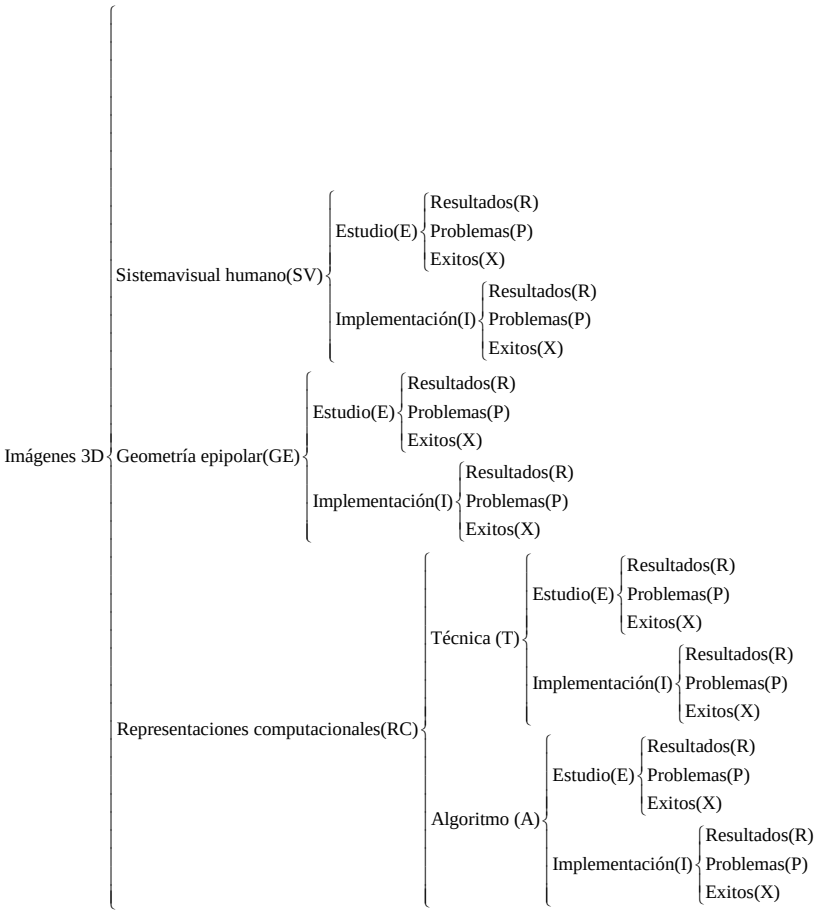


Figura 1.4.- Ventana principal del programa JabRef.

Una vez se recopilaron estos artículos completos con ayuda del programa JabRef versión 2.3.1 (figura 1.4) se clasificaron de la siguiente forma:



Por ejemplo si hasta el momento ya se habían revisado 18 artículos , el artículo número 19 el cual trata sobre el Sistema Visual Humano, es de estudio con resultados experimentales se clasificó como: SVER19, en un campo propio con una etiqueta llamada CLAS.

Capítulo 2

Imágenes Naturales Tridimensionales

En el presente capítulo se describirán las fases del proceso de obtención de una imagen natural tridimensional tomando como modelo al sistema visual humano (véase Anexo A). Como se observa en la figura 2.1, dicho proceso consta de las fases de captura, estimación de profundidad, codificación y visualización.

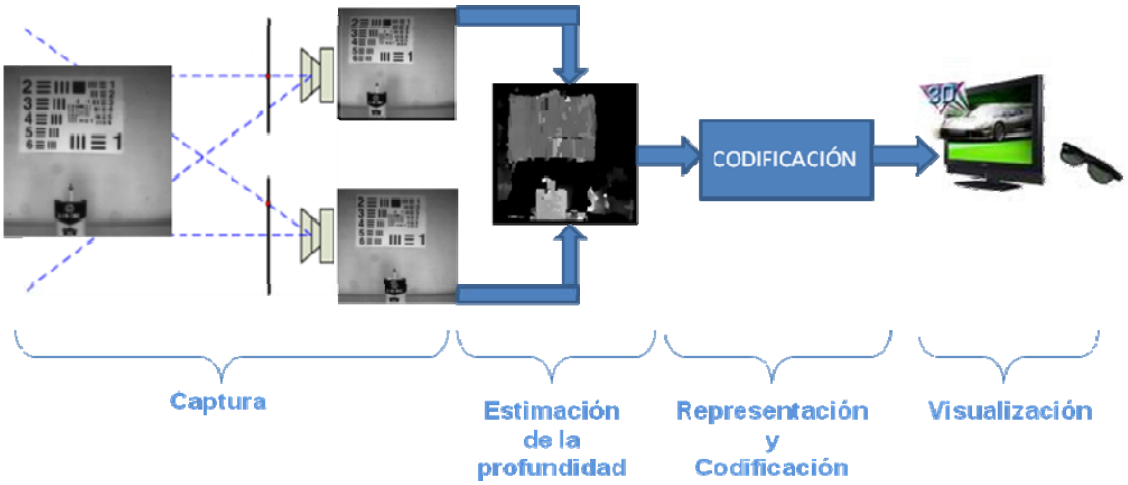


Figura 2.1.- Proceso de obtención de una imagen natural tridimensional.

En la fase de captura se expondrá brevemente cómo se calibran y colocan las cámaras, ya sea de frente emulando la vista humana o hacia el objeto, además de la utilización de diferentes sistemas de cámaras [IRP02]. Una vez que la cámara está en la posición deseada se procede a la obtención de una imagen, dos imágenes o múltiples imágenes [OA05].

Una vez obtenidas las imágenes se pasa a la fase de estimación de profundidad donde se puede trabajar correlacionando un pixel en particular en una imagen con su correspondiente pixel en la otra imagen o también trasladando la información del pixel puro a un apartado de características, donde estas pueden ser reconocidas posteriormente [SB05, ESY08].

Para la fase de representación y codificación existen algoritmos que calculan la disparidad de los pixeles de dos imágenes y en base a ello codifican las diferencias de las imágenes pero también hay los que utilizan la transformada wavelet tridimensional de una imagen. Existen etapas intermedias entre las fases de codificación y visualización, que son las etapas de transmisión y decodificación pero en este trabajo de investigación no se tocarán. En la etapa de visualización se representa de manera en la que el usuario percibirá la estimación de la profundidad, por medio de la auto-estereoscopía o también con la ayuda de gafas polarizadas o HDM's.

2.1 Captura

2.1.1 Calibración y ubicación de cámaras y objetivos

Cuando se requieren ubicar cámaras para capturar imágenes naturales tridimensionales se deberán tomar en cuenta las bases de la geometría epipolar (véase Anexo B), las cuales permitirán clarificar la información que es necesaria para llevar a cabo la búsqueda de correspondientes características a lo largo de líneas [JAP05].

La figura 2.2 ilustra un modelo paralelo, que simula el sistema visual humano, el cual muestra la perspectiva de las dos vistas diferentes de un mismo punto P de un objeto distante, desde los centros de dos cámaras iguales (Fl y Fr), las cuales están separadas sólo en la dirección X por una distancia base [RE06]. En ella también se muestra un par conjugado, es decir, las perspectivas izquierda y derecha del punto P, ilustradas como los puntos Pl y Pr. Al plano resultante de triangular los centros de las cámaras y un punto del objeto en la escena es llamado el plano epipolar. La intersección del plano epipolar con el plano de la imagen es llamada línea epipolar. Según la geometría epipolar, las correspondencias en los putos Pl y Pr deben situarse sobre dicha línea.

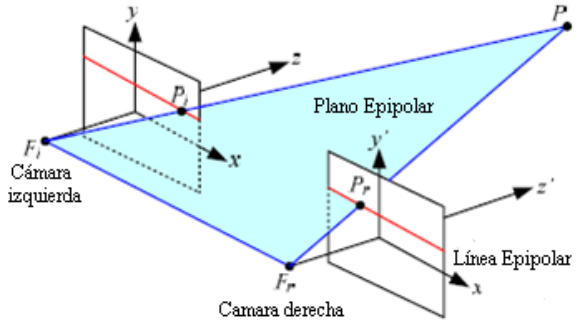


Figura 2.2.- Modelo paralelo.

A diferencia del modelo anterior, el modelo convergente ubica las cámaras no de manera paralela separada solamente por el eje X, sino las lentes se sitúan hacia el objeto por lo que la perspectiva del punto P puede no aparecer en alguna de las imágenes. En la figura 2.3 se observa dicho modelo, el cual en lugar de seguir simples líneas se trabaja con una circunferencia, posicionando al objeto como centro de la misma y, alrededor, las cámaras que lo enfocan.

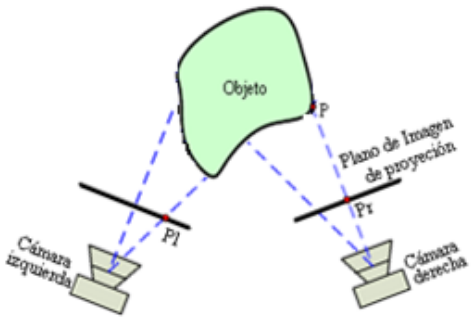


Figura 2.3.- Modelo convergente.

Tanto para el modelo paralelo, como para el convergente, existe la adaptación para múltiples cámaras. Aunque se ha diseñado un modelo de cámara única, figura 2.4, en el cual se colocan frente al objeto dos espejos con una inclinación de 45° con respecto al eje X, al fotografiar ambos espejos se obtendrán dos perspectivas del objeto [SCM91].

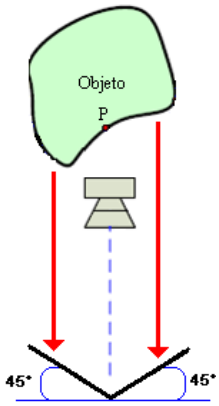


Figura 2.4.- Modelo de cámara única.

2.1.2 Obtención de imágenes

Una vez colocadas las cámaras en las posiciones deseadas, se procede a la obtención de imágenes tridimensionales. Para ello es necesario para ambos modelos, paralelo y convergente, la utilización de dos cámaras que tomarán imágenes de un mismo objeto en diferente perspectiva. Dichas imágenes se consideran bidimensionales sobre el plano de proyección que representa dicho objeto desde la vista de las cámaras. Sin embargo para un modelo de cámara única, el ordenador recibe una única imagen que posee dos perspectivas del mismo objeto, la cual es divida para obtener dos imágenes, como ocurre en los otros modelos. Estas dos imágenes contienen cierta información relacionada que las complementan, que permite obtener la profundidad del objeto o la tercera dimensión de esas imágenes bidimensionales planas. Entonces es determinada la distancia del objeto y su profundidad usando cámaras estéreo [CWR07].

Con las dos imágenes estéreo, un ordenador puede compararlas y concluir que existen partes que son similares y partes que son equivalentes. La cantidad de cambio es llamada disparidad, la cual está relacionada con la distancia del objeto. Mientras más grande es la disparidad del pixel significa que se está más cerca de las cámaras, por el contrario menos disparidad significa que el objeto es más lejano a las cámaras. Entonces, si el objeto es demasiado lejano, la disparidad es cero lo cual significa que el pixel tratado sobre la imagen izquierda es el mismo que el localizado en la imagen derecha.

La figura 2.5 describe las bases geométricas para imágenes naturales tridimensionales usando dos cámaras idénticas. Dichas cámaras son puestas en el mismo plano y dirección, es decir en modelo paralelo. La posición de ambas cámaras es diferente en el eje abscisas y de frente a estas, por conveniencia, se presenta el esquema de proyección de los planos de las imágenes. El punto P del objeto tiene una proyección sobre el plano de imagen izquierdo llamado Pl y otra proyección sobre el plano derecho o llamado Pr. Dichas proyecciones son construidas mediante el trazado de líneas rectas desde el punto P al centro del las lentes de las cámaras. La intersección de dicha línea y el plano de imagen es el punto de proyección.

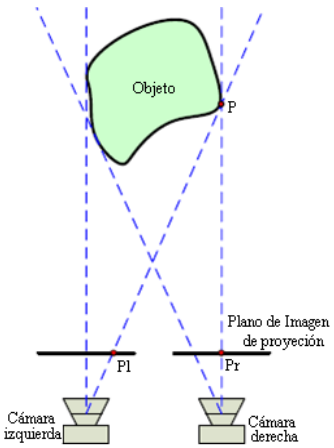


Figura 2.5.- Las posiciones de las dos cámaras y sus respectivos planos de proyección.

También en la figura 2.5 se observa que el punto de proyección de la cámara izquierda Pl, sale del centro hacia uno de los costados del plano de la imagen, mientras que el punto de proyección de la cámara derecha Pr es desplazado al centro. Este desplazamiento respecto al punto correspondiente de la cámara izquierda es utilizado para obtener la información de la profundidad del punto P del objeto observado.

2.2 Estimación de la profundidad

Como se ha expuesto anteriormente, para obtener la información de la profundidad de una escena se necesitan múltiples perspectivas de la misma, por lo regular se procesan un par estéreo [CSS02]. La mayoría de algoritmos y técnicas existentes, intentan igualar ciertas partes de las imágenes. Las diferencias o disparidades, en partes iguales de dos imágenes, es inversamente proporcional a su profundidad en la escena [CCK07]. Esta disparidad puede ser convertida en la profundidad absoluta la cual se denota por la ecuación 2.1.

$$z = \frac{s \times f}{d} \qquad \text{……...…….……..(2.1)}$$

Donde: z, medida en metros, es la distancia entre las cámaras y el objeto.
s, medida en metros, es la separación entre las cámaras.
f, medida en pixeles, es la distancia focal de las lentes de las cámaras.
d, medida en pixeles, es la disparidad calculada.

Dado que la resolución de profundidad es directamente proporcional a la resolución de la imagen, mientras más grande sea esta, mejor precisión se obtendrá, ya que la información de disparidad crecerá también. Resulta poco útil estimar la disparidad absoluta de toda una escena, por que se necesitarán estimaciones parciales para trazar un mapa de profundidad, además generalmente una escena incluye objetos más cercanos y objetos más lejanos, e incluso un mismo objeto puede tener partes más cercanas a las cámaras que otras. En un mapa de este tipo se detectan las distancias respectivas de los objetos de la escena utilizando solamente datos de disparidad, normalmente se representa con una imagen en escala de grises, la cual representa la profundidad en siluetas, siendo las más cercanas representadas con un color blanco y las más lejanas con un color negro [PL07]. Para agrupar los algoritmos que extraen la profundidad de partes de diferentes imágenes, se consideran dos métodos: los basados en área [MK99, STS07] y los basados en características [Mig00].

2.2.1 Basada en área

Básicamente este método trabaja correlacionando un pixel en particular en una imagen con su pixel correspondiente de la otra imagen. Dado que un pixel en una imagen puede ser igualado a muchos pixeles de la otra imagen, los pixeles vecinos son usados para ayudar a encontrar el pixel adecuado. A este conjunto de pixeles cercanos se le conoce como una imagen falsa, la cual es correlacionada con su par falso en la otra imagen, ya con ello es calculada entonces la información de la disparidad de los pixeles. Todo este proceso es repetido en cada pixel de la imagen. Por esta razón, este método es también llamado un esquema basado en la correlación.

El proceso de segmentación de área, o también llamado umbralización adaptativa, permite obtener resultados con un tiempo de cómputo muy bajo. Cada región extraída debe ser caracterizada mediante un vector de características de bloque basadas en la posición, el tamaño, el color y la forma, el cual permitirá la identificación eficiente y particular del objeto [MK99]. El proceso de correspondencia utiliza este vector para emparejar las regiones en base a la similitud que presentan. Esta medida se obtiene mediante la ponderación de las características que forman el vector. Posteriormente se realiza el cálculo de la disparidad y de la profundidad, incorporando un factor de corrección empírico. Por último, en base a la segmentación inicial y a las profundidades calculadas, se detectan los objetos buscados. El modelo presenta ventajas de tiempo de cómputo y de precisión en la estimación de la profundidad y en la detección de objetos.

Para evaluar cuán iguales son los bloques comparados es necesaria una medida de distancia. Algunas de las medidas utilizadas son: la adición de diferencias absolutas (SAD), la adición de diferencias

cuadráticas (SSD) o la correlación transversal. Las restricciones para la estimación de profundidad están normalmente relacionadas con la geometría epipolar.

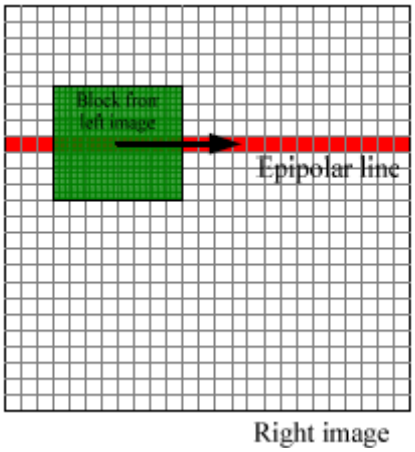


Figura 2.6.- Proyección de un bloque [CWR07].

Normalmente los bloques están definidos por la facilidad de su igualación. Cada bloque en la imagen izquierda es igualado con un bloque en la imagen derecha proyectando el bloque izquierdo sobre un área de pixeles en la imagen derecha, como se muestra en la figura 2.6. En cada desplazamiento, la suma de parámetros comparados, como la intensidad o el color de dos bloques, se realiza y se almacena. La suma de estos parámetros es la fuerza de igualación. El desplazamiento por el cual se obtienen mejores resultados, según los criterios preestablecidos de igualación, es considerado como la mejor correspondencia. El bloque que ha correspondido en las imágenes estéreo es sensible a los cambios de luminosidad, esto se considera ruido en las cámaras izquierda y derecha.

Cuando se trabaja la estimación de profundidad basada en el área normalmente se implementan en los algoritmos que encuentran las coincidencias en las ecuaciones de la adición de diferencias absolutas (2.2), la adición de diferencias cuadráticas (2.3) o la correlación transversal (2.4) [STS07].

$$SAD(p_l, p_r) = \sum_{j=-n}^{n} \sum_{i=-m}^{m} \left| I_l(x_l + i, y_l + j) - I_r(x_r + i, y_r + j) \right| \quad \text{.....(2.2)}$$

$$SSD(p_l, p_r) = \sum_{j=-n}^{n} \sum_{i=-m}^{m} \left(I_l(x_l + i, y_l + j) - I_r(x_r + i, y_r + j) \right)^2 \quad \text{.....(2.3)}$$

$$CORR(p_l, p_r) = \sum_{j=-n}^{n} \sum_{i=-m}^{m} I_l(x_l + i, y_l + j) \times I_r(x_r + i, y_r + j) \quad \text{.....(2.4)}$$

Donde : I_l e I_r son la intensidad de los pixeles en las imágenes derecha e izquierda.

p_l y p_r son la comparación de puntos bidimensionales (x_l, y_l) y (x_r, y_r).

Los puntos (x_l, y_l) y (x_r, y_r) se varían hasta encontrar un valor mínimo en la adición de diferencias absolutas (SAD) y la adición de diferencias cuadráticas (SSD), y un máximo en la correlación transversal. Cuando se estén presentas dichas características en bloque se le considerará a este la mejor correspondencia.

2.2.2 Basada en características

Las técnicas de estimación de profundidad basadas en características inicialmente trasladan los datos del pixel puro a un conjunto de características. Se asume que dichas características son el conjunto de propiedades de la imagen más estables. Con ello, el algoritmo intenta igualar el conjunto de características con las presentes en el correspondiente par estéreo, calculando con dichas características la disparidad de los

pixeles. Si existen pixeles carentes de cualquier propiedad, no tendrán un valor de disparidad asociado [Gri85].

La mayoría de los algoritmos que calculan la profundidad basándose en sus características se apoyan en el algoritmo propuesto Marr y Poggio [MPH79], el cual describe 4 pasos para encontrar una solución al problema de correspondencia estéreo:

1. Tanto la imagen derecha como la izquierda son filtradas con operadores diferenciales orientados, los cuales aumentarán el tamaño cuatro veces y por ende la distancia al centro del plano epipolar. Este filtrado se desarrolla para detectar los cambios significativos en la intensidad en diferentes escalas.

2. Después se localizan en las imágenes los cruces por cero, esto se realiza escaneando en líneas perpendiculares imaginarias a lo largo del operador diferencial. Los cruces por cero marcan donde se encuentran los cambios significativos en la función original de intensidad en diferentes escalas. Con ello son localizadas las terminaciones de líneas y las aristas.

3. Para cada operador de determinado tamaño y orientación, la equiparación sucede cuando se contrastan los signos de las terminaciones de los cruces por cero en las dos imágenes.

4. La información de disparidad es el proceso de igualación, donde se mueve de la utilización de grandes disparidades en zonas de baja resolución a la utilización de pequeñas disparidades en zonas de alta resolución.

Cuando la correspondencia es alcanzada, esta se almacena un búfer dinámico, llamado esquema dimensional 2½.

Después, para obtener el mapa de disparidad inicial, normalmente se decide utilizar un algoritmo tradicional de programación dinámica, como el de Cox [CHR96]. Este algoritmo especifica como parámetro de entrada la varianza estimada del conjunto de valores de intensidades de píxeles de la imagen, por ello se aplica la misma varianza sobre todos los puntos a equiparar. Ello implica la suposición de que todas las zonas de la imagen tienen el mismo grado de uniformidad, sin cuestionar si en la imagen concreta es así o no. Existiendo algunas modificaciones, como la realizada por Satorre [SCB03], donde se recálcula la varianza para cada punto a tratar. Puesto que las igualaciones están localizadas en zonas y siempre dentro de un rango de búsqueda establecido por la restricción de línea epipolar, se calcula la varianza para cada uno de esos rangos. De este modo se ajusta automáticamente el valor de la varianza de los puntos a equiparar. Con esta modificación se consiguen unos costes de oclusión y de emparejamiento que mejoran algunos de los algoritmos existentes, especificando para cada uno de los intervalos de búsqueda, y no para los generales de toda la imagen, lo que aporta a las igualaciones una mayor fiabilidad.

Una vez obtenido el primer mapa de disparidad en el nivel de escalado mayor, la cima de la pirámide, se calculan de modo distinto las igualaciones de los restantes niveles de la pirámide (figura 2.7), aportando la información obtenida en el nivel anterior. Con ella, la función de coste (Ecuación 2.5) calcula las posibles igualaciones, manejando la información de las correspondencias obtenidas en el nivel anterior de la pirámide, junto a la información de aristas, puntos esquina y las posibles igualaciones entre las imágenes izquierda y derecha escaladas al nivel actual de tratamiento [CSR03].

$$Cf_{l-1} = c_{ij,d} - \frac{\left|d_{ij,l} - d_{ij,l-1}\right|}{2DM_l} - \lambda\left(1 - Arista_H\left(L,R\right) - Arista_V\left(L,R\right) - Arista_C\left(L,R\right)\right) \quad \text{.....….(2.5)}$$

Donde: $c_{ij,d}$ es el coeficiente de correlación descrito en la ecuación 2.6.

$$c_{ij,d} = \frac{\operatorname{cov}\left[I_l\left(i,j\right), I_r\left(i,j+d\right)\right]}{\sqrt{\operatorname{var}_{ij}\left[I_l\left(i,j\right)\right]} \times \sqrt{\operatorname{var}_{ij,d}\left[I_r\left(i,j+d\right)\right]}} \qquad \text{……..(2.6)}$$

λ es el coeficiente de regularización sobre la detección de aristas verticales, horizontales y puntos esquina.

$d_{ij,l-1}$ es la disparidad de *p=(i,j)* y se define en la ecuación 2.7.

$$d_{ij,l-1} = \max_{d=SR} \{Cf_{l-1}\} \qquad \text{……..(2.7)}$$

con Cf_{l-1} la función de coste en el nivel *l-1* y DM_l la máxima disparidad en el nivel *l.*

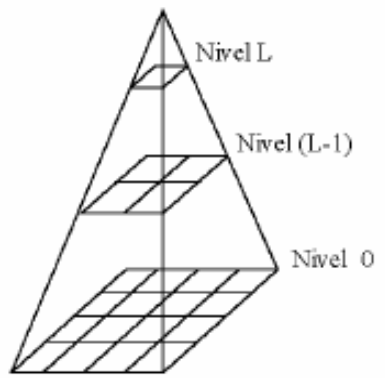


Figura 2.7.- Niveles en una pirámide de disparidad [CSR03].

2.3 Representación y codificación

La percepción de la profundidad dada por sistemas de imagen o vídeo estereoscópico, le pueden ofrecer cierto realismo a muchas aplicaciones como películas en tercera dimensión, cirugías médicas, videoconferencias, aplicaciones multimedia, operaciones controladas a distancia, entre muchas otras. Sin embargo, existen diversos retos a imponerse antes del uso casi sin restricciones de sistemas estereoscópicos, uno de estos retos a vencer es el limitado ancho de banda. Siendo este último el principal obstáculo para que dichos sistemas estéreo sean viables, ya que para enviar, por un canal o almacenar en disco duro, este tipo de imágenes se requiere de al menos el doble de datos que los que se enviarían o almacenarían en un sistema de codificación monocular convencional.

Las técnicas utilizadas para codificar imágenes estéreo son similares que las utilizadas para codificar imágenes monoculares. En general los sistemas estereoscópicos se aprovechan la redundancia entre pares estéreo, además de la redundancia temporal entre tramas consecutivas de cada vista de la escena tridimensional [SCM91]. Por ello, es importante explicar algunas técnicas existentes de codificación de imágenes estéreo. Existen métodos, como el de Duarte en [DCS02], el cual se vale de igualaciones de ciertos patrones recurrentes en las imágenes, además segmentar a dichas imagen en bloques variables, donde cada segmento se contrae, expande o desplaza de acuerdo a los criterios de un diccionario previamente establecido. Otros métodos, como los de Xu en [XXL02] y Nayan en [NEB02], proponen algoritmos para codificar imágenes estéreo utilizando la transformada wavelet, manejando algunas propiedades contenidas en JPEG2000[SMT06, Ols08].

2.3.1 Basada en Patrones recurrentes multiescalares

El método de Duarte expuesto en [DCS02], se trabaja con una imagen referencia, la cual puede ser la izquierda o derecha, la cual debe ser codificada por un método conocido, jpeg por ejemplo. Después se estima el mapa de profundidad de ambas vistas, con ello se obtiene un error denominado Diferencia Compensada de Disparidad o DCD, el cual se codifica para su posterior transmisión o almacenamiento. La calidad del mapa de disparidad provoca que se incremente en gran medida tanto la cantidad de información

que se lleva a la DCD como el número de bits para codificarlo. Por ello alcanzar mapas de disparidad precisos requieren también de grandes tasas de transmisión o almacenamiento.

Una solución a lo anteriormente descrito es el uso de estimadores de disparidad basados en bloques, los cuales pueden producir mapas de disparidad no muy correctos, ya que los bloques no son constantes en cuanto a sus características. Por lo tanto, la estimación de disparidad en bloques jerárquicos es una solución a la cual se le puede sacar provecho. Se usan los bloques de mayores tamaños en zonas donde las variaciones de disparidad son leves, mientras que los bloques pequeños se utilizaran donde exista una mayor variación. Como ya se sabe una imagen estéreo involucra dos imágenes muy semejantes, en donde se pueden aprender los patrones que hay en las mismas.

Este trabajo emplea una nueva clase de codificadores usando patrones recurrentes multiescalares. Un analizador sintáctico multiescalar y multidimensional (MMP) utiliza contracciones y expansiones de elementos pertenecientes a un diccionario para codificar cada segmento de una imagen. Esta imagen es segmentada de acuerdo a un criterio en la tasa de distorsión, entonces el diccionario es actualizado con la concatenación de elementos previamente codificados.

Este método es eficiente para codificación de imágenes monoculares, especialmente cuando la imagen está compuesta por ilustraciones y texto. Con todo lo anterior el presente método propone un codificador basado en un MMP adaptado a imágenes estéreo. Además son también utilizadas las expansiones y contracciones de los elementos en el diccionario, los desplazamientos de los bloques previamente codificados.

En un MMP, la imagen inicialmente se divide en bloques de tamaño $N \times N$. Después cada bloque es segmentado también en bloques más pequeños de tamaños variables. Estos pequeños bloques son aproximados por versiones contraídas o expandidas de matrices en un diccionario. Estas contracciones y expansiones son desarrolladas utilizando un procedimiento similar a las operaciones de interpolación. La salida en un MMP está compuesta de los correspondientes índices del diccionario de cada bloque, de igual modo de la información estimada de las segmentaciones respectivas.

La segmentación en un MMP puede ser representada por un árbol binario. Como ejemplo se tiene la figura 2.8. El nodo raíz del árbol corresponde a un elemento de dimensión $N \times N$. Este elemento puede ser dividido en dos elementos más pequeños (hijos), representados por los dos nodos debajo del nodo raíz. La división, en mitades, puede ser horizontal, como en la citada figura 2.8, o vertical.

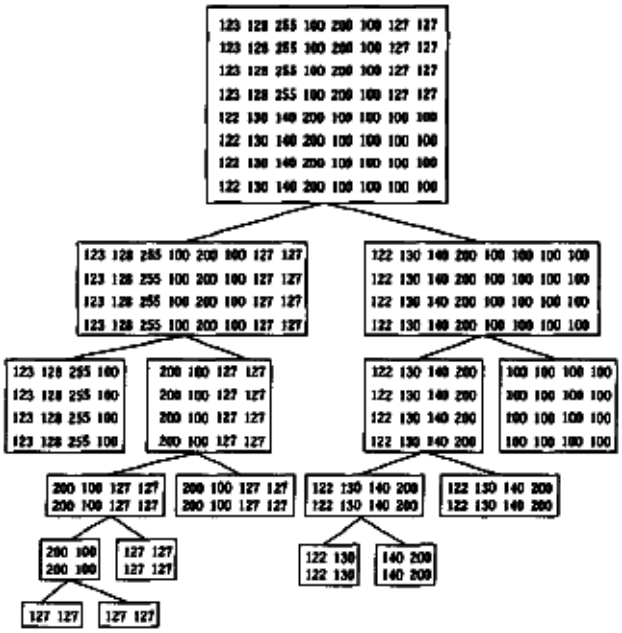


Figura 2.8.- Árbol de segmentación binario

El propósito de un MMP estéreo es codificar directamente los pares estéreo, compuestos por una imagen referencia y una imagen objetivo, sin evaluación explicita de la DCD.

Esta adaptación al MMP es basada en dos puntos principales:

1. La inclusión en el diccionario de elementos correspondientes a los desplazamientos de bloques previamente codificados de la imagen referencia.
2. La utilización de bloques de tamaños variables, una propiedad básica de un MMP. Si el coste de utilización de un bloque grande es mayor que el coste de un bloque pequeño, la segmentación se ejecuta. Así, los desplazamientos pueden ser representados con la precisión que se requiriese.

Inicialmente, tanto la imagen referencia como la objetivo son subdivididas en $N \times N$ bloques. Los bloques adyacentes son agrupados en forma de fila de bloques, también conocida como una ROB, es decir, una ROB es un trozo $N \times M$ de la imagen, donde M es el número de columnas que posee el bloque. Cada bloque de la primera ROB de la imagen referencia es procesada por un MMP clásico. Después de ello, el diccionario del MMP obtiene las aproximaciones a todos los $N \times N$ bloques, de igual modo las aproximaciones a todas los sub-bloques que resulten del proceso de segmentación.

Asumiendo que la orientación de las cámaras es paralela, un $n \times m$ bloque en la primera ROB de la imagen objetivo corresponderá con otro $n \times m$ bloque en la primera ROB de la imagen referencia, pero desplazado en su posición. El desarrollo de un MMP mejora la igualdad de bloques, ya que cuando esto se implementa existe una mayor frecuencia en encontrarla. Así, para incrementar la probabilidad en la igualación, se incluyen versiones desplazadas de los bloques en el diccionario, obtenidas después del procesamiento de la primera ROB de la imagen objetivo. La adición de bloques desplazados es repetida, hasta alcanzar la última ROB de la imagen referencia, pero siempre codificando antes la respectiva ROB de la imagen objetivo.

En un MMP estéreo, la inclusión de elementos desplazados en el diccionario reemplaza los procedimientos de estimación de profundidad. Cuando la opción de un elemento resulta costosa, el proceso de decisión, el cual especifica el mejor árbol de segmentación, concluye que se debe dividir el bloque en dos partes. Esto es equivalente, en los métodos basados en la DCD, para obtener las disparidad utilizando bloques cada vez más pequeños, alcanzando mayor precisión en las disparidades estimadas. Los elementos correspondientes al desplazamiento equivalente a medio pixel, también son incluidos para mejorar la resolución de la estimación.

Bajo condiciones donde las cámaras son posicionadas de manera paralela hacia el objeto, se asume que las disparidades se encuentran solamente de manera horizontal, con ello se reduce el número de elementos desplazados. Además este número es limitado por una búsqueda en alguna ventana dada por valores de disparidad, ya sean mínimos o máximos, en la imagen objetivo. Los estadígrafos, utilizados en los elementos del diccionario y que además están agrupados por tipo, muestran que cuando la imagen objetivo es codificada, los elementos desplazados son mucho más utilizados que los que no que no lo están. Entonces el diccionario es dividido en dos partes: la primera contiene los elementos desplazados y la segunda que contiene todos aquellos elementos que no lo están. Una bandera es usada para señalar cual diccionario está siendo seleccionado.

2.3.2 Basadas en la Transformada Wavelet

Esta sección se basa de los trabajos expuestos por Xu en [XXL02] y por Nayan en [NEB02], donde se describen dos diferentes algoritmos para codificar imágenes estéreo utilizando la transformada discreta wavelet (DWT) [BS00, SR03]. En la propuesta de Xu se describe un algoritmo, el cual se basa en el diagrama presentado en la figura 2.9, que como primer paso se procede a aplicar la transformada wavelet en la imagen izquierda, después se aplica una codificación wavelet Embedded [BS02, JZS02] a los coeficientes obtenidos. Con el fin de correlacionar dicha imagen izquierda con la imagen derecha, y así estimar y compensar la disparidad, se reconstruye a la imagen izquierda decodificándola y realizándole una transformación wavelet inversa. Lo resultante de este contraste es codificado de igual manera como se codificó la imagen izquierda original.

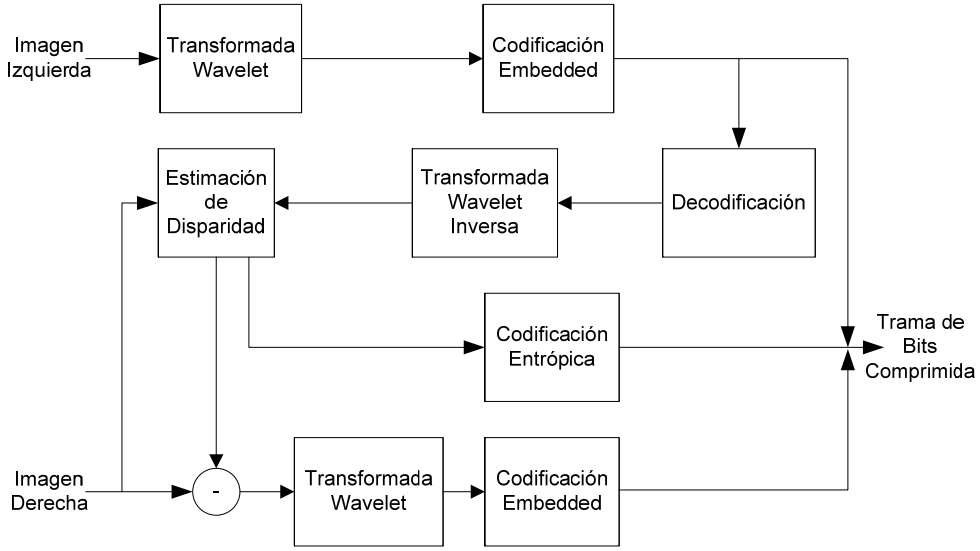


Figura 2.9.- Diagrama de codificación wavelet de imágenes estéreo.

Tanto la imagen izquierda como la disparidad residual compensada de la imagen derecha son convertidas al dominio wavelet utilizando tres niveles de esta (figura 2.10). Se codifica cada sub-banda de una imagen usando un codificador aritmético adaptativo basado en contextos. De igual manera que la EBCOT en JPEG2000 [SMT06, Ols08], los coeficientes wavelet en cada sub-banda son codificados plano de bits por plano de bits, y por cada uno existen tres pasos:

1. Propagación de significancia, los coeficientes son codificados aunque no sean significativos, mientras que tengan vecinos que si lo sean. La información sobre si un coeficiente es significativo o no en el actual plano de bits es codificado utilizando codificación aritmética basada en el contexto. Esta operación es llamada codificación del cero. Si el coeficiente se convierte en significativo, su signo es también enviado al codificador aritmético basado en contexto utilizando un modulo de codificación de signo.

2. Refinamiento de magnitud, los coeficientes son codificados cuando estos son significativos en el anterior plano de bits. Para esos coeficientes, sus bits significativos y sus signos se codifican también. Así en el paso de refinamiento, los bits del actual plano de bits son agregados para afinar las magnitudes de dichos coeficientes, donde también se utiliza el codificador aritmético.

3. Normalización, durante este paso, los coeficientes faltantes son codificados, aunque no lo hayan sido en pasos anteriores. Estos coeficientes no son significativos, por lo cual tanto la codificación del cero como la del signo son aplicadas en este paso.

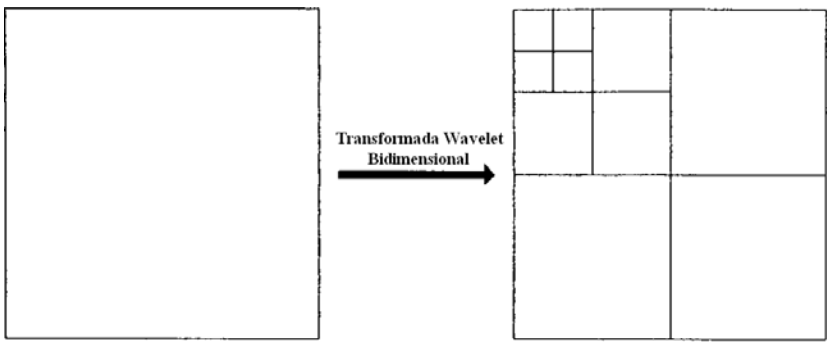


Figura 2.10.- Tres niveles de transformada Wavelet aplicados a una imagen bidimensional.

El algoritmo utiliza las mismas asignaciones para codificación de cero, codificación de signo y el refinamiento de magnitud que utiliza JPEG2000 para un codificador aritmético basado en contexto. Una diferencia entre el presente esquema y EBCOT en JPEG2000 es que no se dividen la sub-bandas en bloques. La ventaja dividir las sub-bandas en bloques en JPEG200 es alcanzar una optima tasa de distorsión entre bloques. Otra ventaja más es que se puede tener un acceso aleatorio a los bloques. Sin embargo codificar una sub-banda como un todo en este esquema tiene las siguientes ventajas:

- El tamaño de una imagen estéreo no es tan grande y las propiedades estadísticas de los coeficientes wavelet son similares dentro de una sub-banda. Inclusive si se dividen las sub-bandas en bloques y se codifican de forma independiente, la ganancia de codificación es circunstancial si la asignación de bits es óptima. Por el contrario una codificación de bloque independiente, reduce el número de muestras de preparación para la adaptación del contexto, causando un problema de disolución del mismo y afectando al desarrollo de la propia codificación.

- Al codificar cada sub-banda como un todo, no se representarán muchos bits innecesarios de información principal de los bloques, entonces no será necesario codificar la asignación de bits o alguna otra información de cada bloque. Con ello se podrá tomar solamente una parte de la trama de bits resultante. Esto es especialmente útil cuando se tiene el caso de una tasa de bits baja.

Nayan en su propuesta genera los coeficientes aplicando una cascada de bancos filtrados de dos canales a la imagen. La figura 2.11 representa una descomposición en tres niveles de transformada wavelet de una imagen.

Normalmente en un proceso de codificación de sub-bandas de una transformada wavelet discreta, los coeficientes son agrupados en conjuntos orientados por sub-bandas (sub-banda común, localización espacial diferente) mientras que en los procedimientos de codificación de sub-bandas en la transformada discreta del coseno, se agrupan en bloques (localización espacial común, sub-bandas diferentes).

La idea detrás de los bloques wavelet es agrupar los coeficientes wavelet en bloques como se muestra en la citada figura 2.11, así el agrupamiento es similar al utilizado en los procedimientos de codificación de sub-bandas con la transformada discreta del coseno.

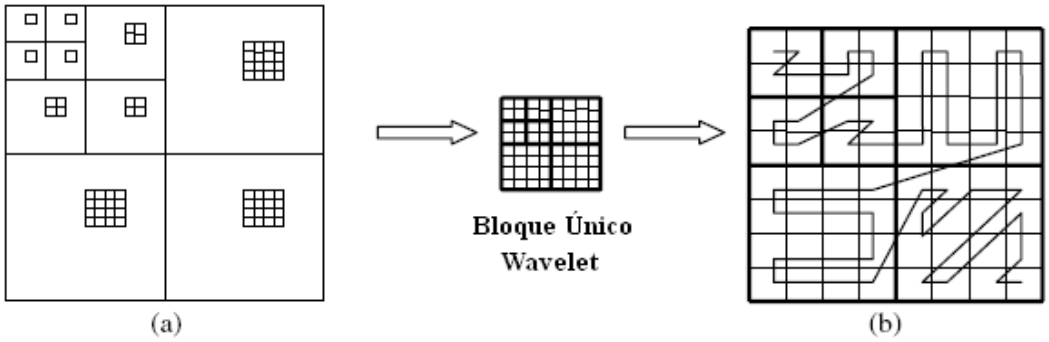


Figura 2.11.- Formación de un bloque único wavelet desde un tercer nivel de descomposición.
(a) Imagen descompuesta. (b) Orden de escaneo por un bloque único wavelet

Las figuras 2.12 y 2.13 representan el diagrama de bloques del codificador y decodificar (CODEC), respectivamente, propuesto por Nayan. Se observa que se asume que los valores de disparidad son satisfactoriamente codificados y trasmitidos o almacenados sin perdidas al decodificador. En el codificador, la imagen derecha es dividida en bloques de pixeles no traslapados de 8×8.

En cada bloque, se desarrollada una búsqueda en la imagen izquierda original dentro de una área con la máxima probabilidad, considerando a esta la imagen referencia, hasta encontrar la mejor predicción. Asumiendo que se utiliza la posición paralela de las cámaras para obtener el par estéreo, lo que limitaría a

que la ventana o área de búsqueda solamente se realizara en dirección horizontal, extendiéndose a la izquierda hasta el bloque corresponda.

En el procedimiento de búsqueda citado, se encuentra la mejor predicción utilizando el error cuadrático medio o MSE como criterio de predicción. La predicción de errores de todos los bloques de la trama derecha no traslapados son utilizados para formar la imagen con la predicción del error o imagen PE. La imagen izquierda, L y la imagen PE sufren una descomposición de tres niveles en la transformada discreta Wavelet separadamente y son subsecuentemente convertidas en sus representaciones de bloque wavelet.

Cada bloque wavelet de dos representaciones entonces experimentan un rastreo como se muestra en la figura 2.11b a diferencia del procedimiento de escaneo en zigzag adaptado por JPEG. Esta modificación en el escaneo es básica para la representación de bloques wavelet, dado un bloque, existen coeficientes, los cuales no sólo pertenecen a diferentes sub-bandas, sino también a coeficientes múltiples, los cuales pertenecen a la misma sub-banda. Después del procedimiento de escaneo modificado discutido anteriormente, los coeficientes ordenados sufren una cuantificación escalar.

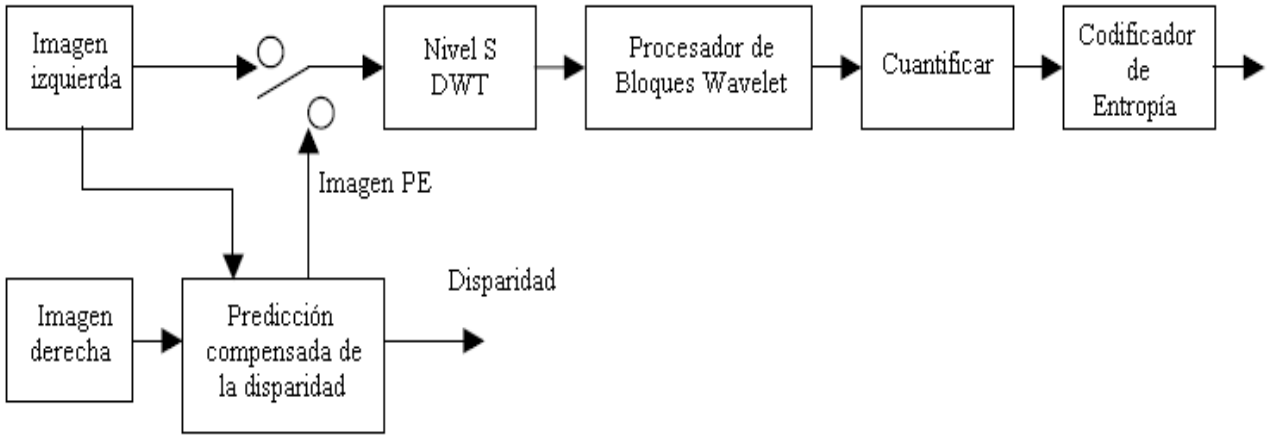


Figura 2.12.- Codificador.

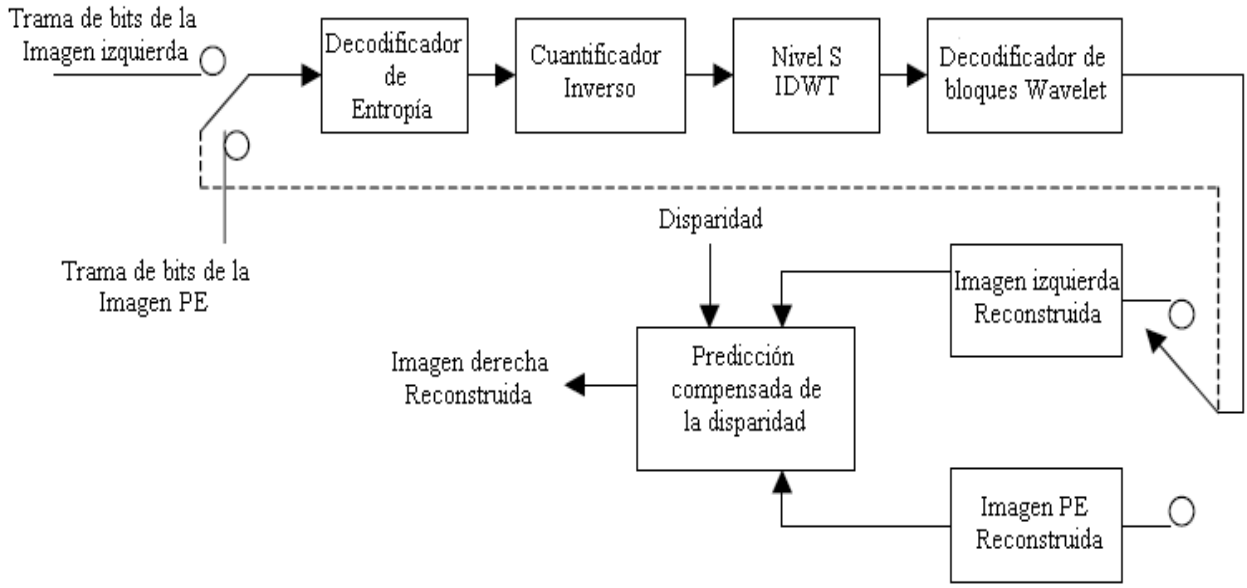


Figura 2.13.- Decodificador.

2.4 Visualización

Como se ha visto hasta el momento existe una amplia variedad de técnicas de obtención, estimación y codificación de la profundidad, aunque cuando se presentan sus diferentes resultados todas expresan dichos resultados en únicamente dos variantes: las autoestereoscópicas y las que utilizan gafas para polarizar la luz.

La forma tradicional de cómo trabajan los sistemas estereoscópicos es empleando ciertos aparatos los cuales dirigen las imágenes derecha e izquierda directamente al ojo indicado. Ejemplo de esto son algunos cascos de realidad virtual o HMD's o gafas obturadoras. Para los estudiosos del tema de imágenes en tercera dimensión, cualquiera de los sistemas en 3D debería simular en su totalidad a la visión humana, es decir, debería ser autoestereoscópica, ya que así no se requerirían de las citadas gafas o condiciones especiales de iluminación. En la tabla 2.1 se contrastan algunas características de los visores de imágenes naturales en tercera dimensión que existen.

	Espejo estereoscópico plano	**Goggles activos con trama secuencial CRT**	**Pantalla obturadora con trama secuencial CRT**	**Auto estéreo con zona única**	**Auto estéreo con zonas múltiples**	**Anáglifo**	**HMD**	**Doble LCD's apilados**	**Doble Monitor con caminos ópticos aislados**
resolución UXGA o mejor	x					x			x
Nitidez de más de 70 cd/m2 por ojo	x			x	x		x		x
Contraste mayor de 200:1	x						x		x
Cambio de 2D a 3D	x	x	x	x	x			x	
Gafas	polarizadas	Goggles activos	Polarizadas	Ninguna	Ninguna	Rojas-Azules	Goggles activos	polarizadas	Ninguna
Múltiples usuarios	x	x	x		x	x		x	
Libre de parpadeo	x			x	x	x	x	x	x

Tabla 2.1.- Comparación entre tecnologías [WEB11].

2.4.1 Autoestereoscópica

Las pantallas autoestereoscópicas son atractivas porque ofrecen la más cercana aproximación de cómo es el mundo que se encuentra alrededor de las personas, sin trabas como son el uso de artefactos externos de visualización. Si se desean generar las más simples imágenes autoestereoscópicas se tendrá que recurrir a los métodos de pantalla selectora, los cuales incluyen las barreras de paralelaje y las laminas lenticulares, que son muy parecidos entre sí [VMP04].

Los métodos de barreras de paralelaje son conocidos desde comienzos del siglo XX, estos métodos incluyen el paralelaje estereográfico y el relacionado con el paralelaje panorámico. El paralelaje estereográfico consiste en un material opaco con finas divisiones verticales a distancias iguales. Cada división transparente actúa como una ventana a una parte vertical de la imagen localizada detrás de ella, la porción extraída dependerá de la posición del ojo, como se observa en la figura 2.14.

La imagen de paralelaje estereográfico está hecha intercalando las columnas de dos imágenes. En la figura 2.15 se observa el caso en el que se combinan dos imágenes con perspectivas diferentes, la del ojo derecho y la del ojo izquierdo, en pequeñas ranuras verticales intercaladas, las que dan como resultado una

subimagen que contiene ambas perspectivas. Paralelaje panorámico es muy parecido al estereográfico con la diferencia que se combina una cantidad N de imágenes por lo que se tiene una cantidad N de perspectivas a combinar en la imagen resultante [GHB08].

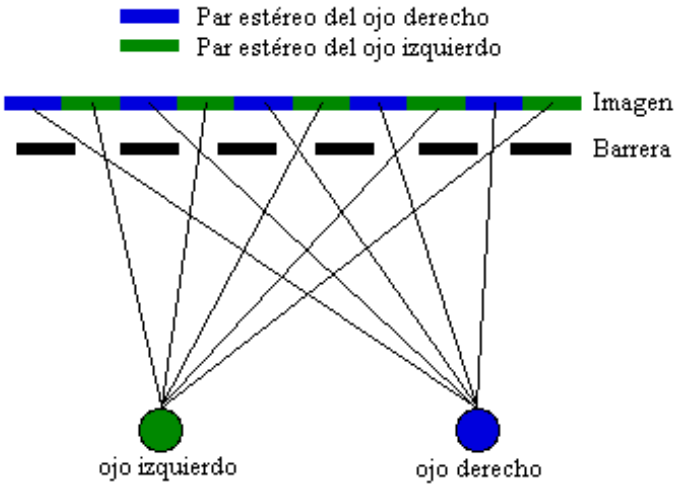


Figura 2.14.- Paralelaje estereográfico.

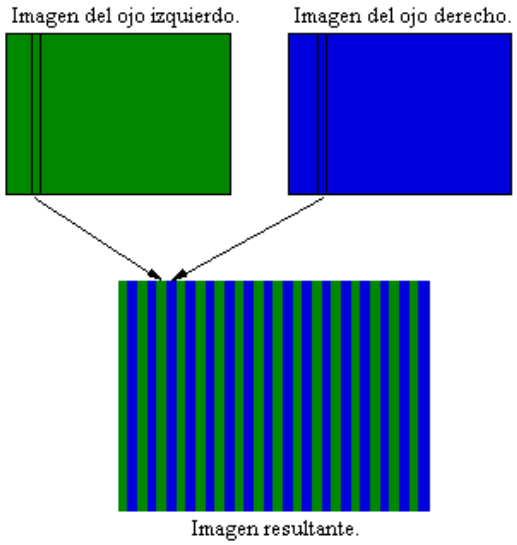


Figura 2.15.- Combinación de perspectivas.

Al igual que el Paralelaje Panorámico las imágenes lenticulares pueden combinar una cantidad N de imágenes. Aunque las imágenes lenticulares utilizan unas hojas, las cuales contienen una serie de lentes cilíndricas, comúnmente, en un substrato plástico.

Las lentes se enfocan a una imagen en la parte de atrás de dicha hoja (figura 2.16). Esta imagen lenticular está desarrollada para enfocar a cada ojo la línea de vista de diferente franja. La razón de localizar la imagen detrás de la hoja lenticular, es para que combine, en un solo punto de la línea de vista, la cantidad N mencionada de imágenes.

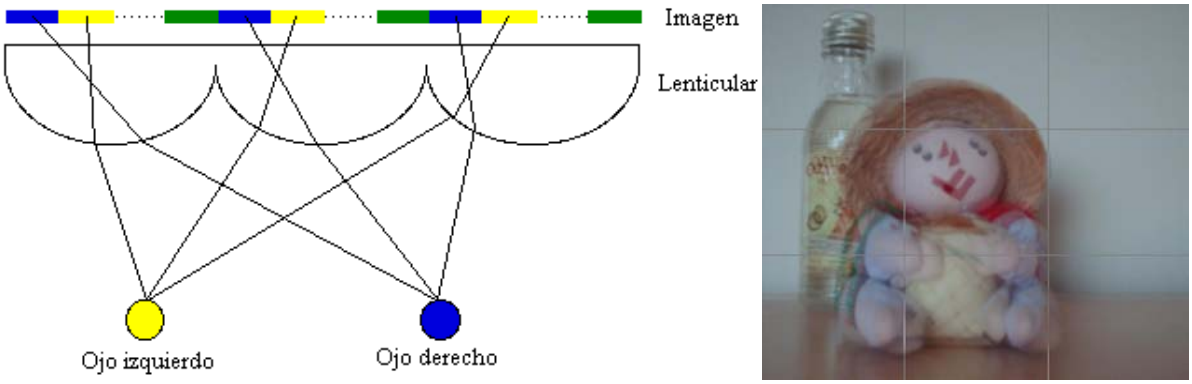


Figura 2.16.- Hoja e Imagen lenticular.

La clave del éxito de la creación de imágenes autoestereoscópicas basadas en hojas lenticulares es la calidad y la uniformidad de las lentes cilíndricas. La hoja normalmente está hecha para que la parte trasera sea exactamente la distancia focal de la imagen, por lo que la información de la imagen surge paralela desde cada lente [ESY08].

Con los principios anteriormente desarrollados algunos investigadores como [VMP04] y [MP04] han realizado pantallas tridimensionales lenticulares, que en lugar de utilizar una imagen fija se colocan proyectores detrás de las lentes cilíndricas, como se puede observar en la figura 2.17.

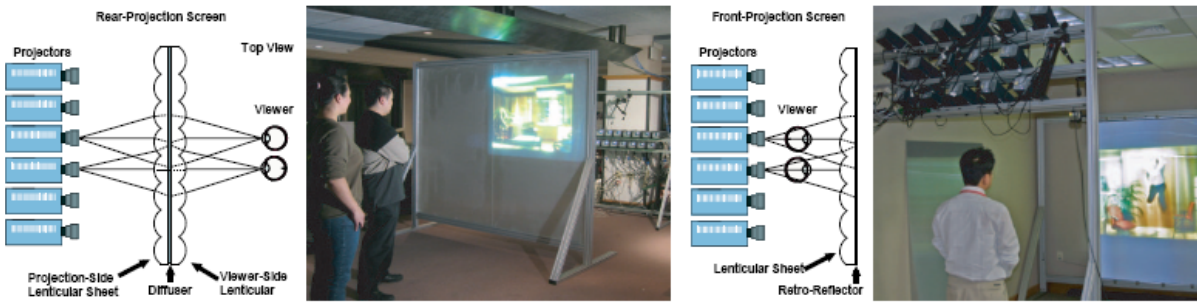


Figura 2.17.- Pantallas tridimensionales lenticulares [VMP04].

2.4.2 Mediante la utilización de gafas

Para la gente nacida entre los años 70's y 80's, asociar la tercera dimensión no se puede separar de los viejos anteojos rojo-azul, mejor llamados anáglifo, aunque para los nacidos en la década de los 90's, probablemente la asociarán con un casco tridimensional o HMD, pero si se les preguntara a la los niños de hoy, seguramente esperarían que la caja de televisión enviara la tercera dimensión por sí sola.

Aunque las gafas anáglifo sean de los 80's (figura 2.18a), su utilización desde entonces no ha dejado de desarrollarse, estas básicamente se utilizan para visualizar imágenes multiplexadas en longitud de onda. En una superficie plana se muestra una imagen a partir de la combinación de dos imágenes desplazadas, creadas únicamente con dos colores complementarios, ya sean rojo-azul, rojo-verde o bien ámbar-azul. Estas dos imágenes equivaldrán al par estéreo.

(a) [WEB12]

(b)

Figura 2.18.- Gafas e imagen anáglifo.

La percepción de profundidad en el sistema visual humano de imágenes en superficies planas requiere la ayuda de experiencias previas o de objetos externos, como pueden ser las gafas anáglifo, gafas LCS (*Liquid Crystal Shutters*, Obturadores de Cristal Líquido) y otros sistemas más modernos, que no provocan cansancio visual. En el caso de las gafas anáglifo, estas están formadas por dos lentes (muy sencillas), cada una con uno de los dos colores que componen la imagen. De esta manera actúan como filtro y dejan ver a cada ojo sólo el par estéreo que le corresponde.

Así pues, por ejemplo, si se tuviera una imagen creada a partir del desplazamiento de una imagen azul (enfocada para el ojo izquierdo) y otra roja (enfocada para el ojo derecho), se necesitarían unas gafas anáglifo con filtros de los mismos colores: el ojo derecho tendría la lente de color azul y el izquierdo la lente roja, ya que el filtro sólo permite ver la imagen que no sea del mismo color. Cabe mencionar que las gafas anáglifo permiten ver en relieve tanto imágenes en papel como en diapositivas o pantallas de cine o de ordenador. Con ellas se crea un efecto satisfactorio, aunque se pierde mucha luminosidad y los filtros utilizados no acaban de conseguir una reconstrucción suficientemente buena en color de la imagen en tercera dimensión.

Las gafas del siglo XXI son simples lentes polarizadas de papel, ya sea polarizado lineal (el más común) o polarizado circular, para usarse con proyectores LCD. Aunque existen las lentes tridimensionales para uso personal, estas son gafas de alta calidad para profesionales. Su calidad óptica permite disfrutar las imágenes tridimensionales con mayor definición, normalmente utilizadas en la visualización de imágenes tridimensionales naturales en pantallas de espejo estereoscópico (figura 2.19). Este tipo de pantallas producen imágenes tridimensionales con resolución WUXGA (1920 x 1200) y que permite que sea utilizado por múltiples usuarios, aunque se sitúen en lugares diferentes. Además, no se necesita una ambientación del entorno para que se perciba la tercera dimensión y gracias a sus pantallas LCD no tiene un barrido horizontal como las CRT.

Con ayuda de dos cámaras separadas a una distancia parecida a la del ojo humano, se capturan las imágenes que por medio de la pantalla LCD inferior, se puede visualizar la imagen izquierda y por el contrario la imagen derecha se ubica en la pantalla LCD superior. A 45° entre las pantallas se coloca un medio espejo polarizador lineal, el cual tiene la función de enfocar a la imagen derecha al ojo derecho y dejar pasar la imagen izquierda sin que sufra ningún cambio. Con ello y la ayuda de las lentes tridimensionales, el cerebro tiene toda la información para formar la tercera dimensión.

Figura 2.19.- Pantallas de espejo estereoscópico [WEB11].

Hay sistemas como el desarrollado por Choi en [CKC04], en el cual se utilizan los antes mencionados Obturadores de Cristal Líquido (Figura 2.20), los cuales alternan su visualización de una pantalla, cuando en esta se está presentando la imagen izquierda, el obturador derecho se polariza inversamente, lo que impide el paso por este y permite el paso de la imagen por el obturador izquierdo. Cada imagen se mantiene por aproximadamente 16 milisegundos antes de cambiar a la imagen del otro ojo, con lo que se visualizarán 30 pares estéreo por segundo.

Figura 2.20.- Obturadores de Cristal Líquido [CKC04].

Capítulo 3

Estimación de la Profundidad

En este capítulo se expondrán cuatro técnicas para encontrar la correspondencia que una imagen izquierda tiene sobre una imagen derecha y con ello poder estimar la profundidad que existe en la escena. Las técnicas de correspondencia son las que se enuncian a continuación:

1. De igualación de bloques [PMF85].
2. De correlación en imágenes no rectificadas [XZ96].
3. Basada en la mejor igualación [TV98].
4. Programación dinámica [CHR96].

Una vez analizadas estas cuatro técnicas, se propondrá una quinta, la cual es la evolución de las mismas, utilizando un Modelo de análisis de varianzas de bloques aleatorizados de dos factores. Inicialmente este modelo fue creado para la investigación agrícola, pero hoy en día también se aplica en otros campos como en la biología y principalmente en experimentos de laboratorio de cualquier índole. Por citar un ejemplo, cuando se aplica un fertilizante a una parcela que tiene diferentes tipos de suelo, se obtiene un resultado expresado como la cantidad en kilogramos de un producto de la cosecha, lo que depende directamente de qué tipo de fertilizante se utilizó y en qué tipo de suelo se sembró. Análogamente, las imágenes son parcelas las cuales se pueden subdividir en bloques de pixeles, que comparados con la intensidad de cada pixel de algún otro bloque de igual tamaño en otra imagen, arrojarán ciertas características que comparten o que los distinguen. Con el análisis de varianzas de bloques aleatorizados se podrá concluir si dichos bloques de pixeles entre sí son:

- Exactos.
- No Exactos, de pixeles diferentes o iguales intensidades pero en posiciones diferentes.
- Similares.
- Homogéneos.
- Heterogéneos.

Además, se realizarán experimentos con resultados para comprobar las conclusiones anteriormente mencionadas.

3.1 Técnicas existentes

3.1.1 Igualación de bloques

A partir de dos imágenes estéreo se puede intentar una igualación por bloques tomando como distancias las intensidades en el color, no las intensidades de un solo pixel, sino las obtenidas en el entorno de vecindad de cada uno, tal y como se muestra en la ecuación 3.1 [PMF85].

$$\sum_{k=0}^{m}\sum_{l=0}^{n} I_l\left(i+k, j+l\right) \quad \text{……..…….……..(3.1)}$$

Se divide la imagen izquierda (M x N) en bloques de tamaño m x n, es decir M/m x N/n bloques. Para que cada bloque de la imagen izquierda sea buscado en el bloque correspondiente de la imagen derecha se debe utilizar alguna medida de similitud, como por ejemplo, el error cuadrático medio (ecuación 3.2).

$$MSE_{(i,j,d)} = \frac{1}{m \times n}\sum_{k=0}^{m}\sum_{l=0}^{n}\left[I_l\left(i+k, j+l\right) - I_r\left(i+k-d, j+l\right)\right]^2 \quad \text{……..(3.2)}$$

siendo *d = disparidad :* $i_l - i_r$. La búsqueda se realiza en un intervalo $\left[0, d_{max}\right]$

i= número de pixel en el eje *x* de la imagen.

La disparidad de bloque es el valor de *d* para el cual el MSE es mínimo. Si hay varios mínimos se asigna como disparidad aquel valor con el que se diferencia menos con algún vecino. Con todo esto se obtiene una matriz de disparidades de bloque.

Una vez que se obtiene la matriz de disparidades de bloque se trata de calcular la disparidad escalar entre todos los puntos de ambas imágenes. Para ello se filtra la matriz de disparidades de bloques mediante un filtro de mediana, posteriormente se calcula la disparidad de cada píxel de la imagen izquierda mediante la disparidad menor de entre los 8 píxeles vecinos al que pertenece el punto. Finalmente se aplica un filtro de mediana a la matriz global de disparidades de puntos.

En la figura 3.1 se observa un ejemplo de un par estéreo del mundo real al que se le aplica la técnica de igualación de bloques.

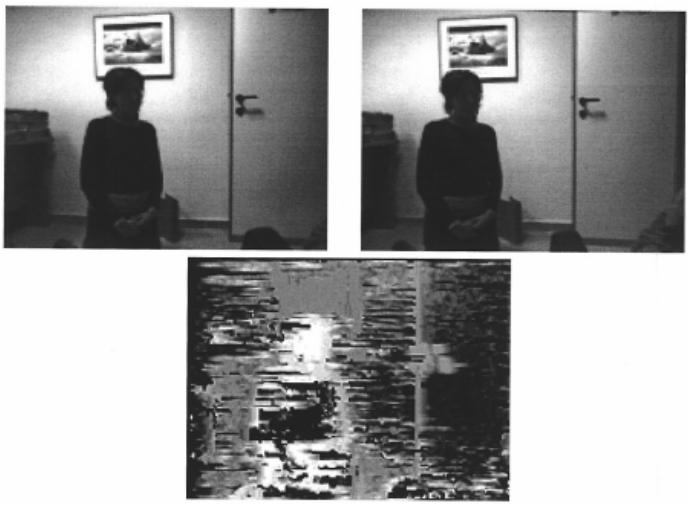


Figura 3.1.- Igualación por bloques.

3.1.2 Correlación en imágenes no rectificadas

Está técnica fue expuesta por Xu en 1996 [XZ96], la cual indica que para encontrar la correspondencia a un punto dado de la imagen izquierda $\left(I_l\left(i_l, j_l\right)\right)$ se ha de realizar una búsqueda en toda la imagen derecha. Dado un punto p_l en la imagen izquierda (I_l), utiliza una ventana de correlación de tamaño $\left(2n+1\right)\times\left(2m+1\right)$ centrado en el punto antes mencionado (figura 3.2). Entonces se selecciona un área de búsqueda rectangular de tamaño $\left(2d_i+1\right)\times\left(2d_j+1\right)$ (donde d_i y d_j son disparidades tomadas a

priori) alrededor de este punto en la imagen derecha, y ejecuta el operador de correlación sobre una ventana dada entre el punto p_l de la imagen izquierda y todos los demás puntos p_r que se encuentran en el área de búsqueda de la imagen derecha, es decir, que no se debe buscar en toda la imagen, sólo en dicha área de exploración. La limitación a una ventana de búsqueda indica algún conocimiento previo sobre las disparidades entre los puntos a igualar. Esto es equivalente a reducir el área de búsqueda para un punto correspondiente de toda la imagen a una ventana dada. Si no se dispone de este conocimiento previo, entonces se debe buscar en la totalidad de la imagen. La correlación se define tal y como se muestra en la ecuación 3.3.

$$\frac{\sum_{u=-n}^{n}\sum_{v=-m}^{m}\left\{\left[I_l(i_l+u,j_l+v)-\overline{I_l(i_l,j_l)}\right]\times\left[I_r(i_r+u,j_r+v)-\overline{I_r(i_r,j_r)}\right]\right\}}{(2n+1)\times(2m+1)\times\sigma_l(i_l,j_l)\times\sigma_r(i_r,j_r)} \qquad \text{……..(3.3)}$$

Donde los valores $\overline{I_l(i_l,j_l)}$ y $\sigma_l(i_l,j_l)$, representan la intensidad media y la desviación estándar en la imagen izquierda en el punto (i_l,j_l), estas se calculan con las expresiones de las ecuaciones 3.4 y 3.5. Los valores para la imagen derecha, $\overline{I_r(i_r,j_r)}$ y $\sigma_r(i_r,j_r)$, se obtienen de forma similar.

$$\overline{I_l(i_l,j_l)}=\sum_{u=-n}^{n}\sum_{v=-m}^{m}\frac{I_l(i_l+u,j_l+v)}{(2n+1)\times(2m+1)} \qquad \text{……..(3.4)}$$

$$\sigma_l(i_l,j_l)=\sqrt{\frac{\sum_{u=-n}^{n}\sum_{v=-m}^{m}\left[I_l(i_l+u,j_l+v)-\overline{I_l(i_l,j_l)}\right]^2}{(2n+1)\times(2m+1)}} \qquad \text{……..(3.5)}$$

El rango de igualación va desde -1, cuando las dos ventanas de correlación no tienen nada similar, a 1, cuando las ventanas correlación son idénticas [WO00].

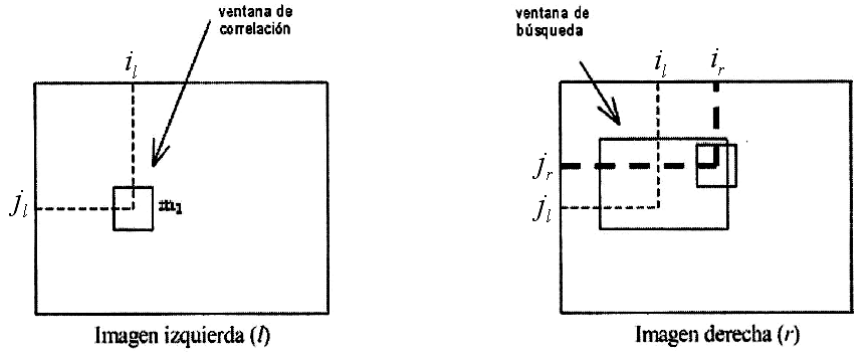


Figura 3.2.- Correlación en imágenes no rectificadas.

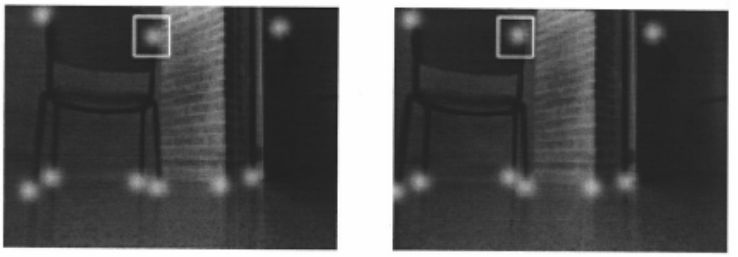
3.1.3 Correlación basada en la mejor igualación

Trucco en [TV98], basa el análisis en el desconocimiento cuantitativo de los parámetros de la cámara, que comienza con las suposiciones generales de muchos métodos para encontrar correspondencias en pares de imágenes. Dichas suposiciones son las siguientes:

- Muchos puntos de la escena son visibles desde ambos puntos de vista.
- Las correspondencias de las regiones de la imagen son similares.

Estas suposiciones son válidas para sistemas estéreo en los que la distancia del plano de proyección de la cámara es mucho mayor que la línea de fondo. En general, ambas suposiciones pueden ser falsas y el problema de la correspondencia llega a ser considerablemente más difícil. A partir de ahora, se tomará la validación de estas suposiciones para garantizar y observar el problema de la correspondencia como un problema de búsqueda: dado un elemento en la imagen izquierda, se busca el elemento correspondiente en la imagen derecha [GCA07]. Esto implica dos decisiones:

- ¿Qué elementos de la imagen igualar?
- ¿Qué medida de similitud adoptar?



(a) Imagen izquierda (b) Imagen derecha

Figura 3.3.- Correspondencia basada en la correlación de pixeles [SCB03].

Se toman como datos de entrada las imágenes de un par estéreo, I_l (imagen izquierda) e I_r (imagen derecha). También se denotan a p_l y p_r como píxeles de la imagen izquierda y derecha, $2W+L$ el ancho (en píxeles) de la ventana de correlación, $R(p_l)$ la región de búsqueda en la imagen izquierda asociada con p_l, y $\psi(p,q)$ una función con dos píxeles (p,q) como parámetros.

Para cada píxel $p_l = [i_l, j_l]^T$ de la imagen izquierda:

1. Para cada desplazamiento $d = [d_1, d_2]^T \in R(p_l)$ se calcula, por medio de la ecuación 3.6.

$$c(d) = \sum_{u=-W}^{W} \sum_{v=-W}^{W} \psi\left[I_l(i_l+u, j_l+v), I_r(i_r+u-d_1, j_r+v-d_2) \right] \quad \text{.....(3.6)}$$

2. La disparidad de p_l es el vector, se representa en la ecuación 3.7.

$$\overline{d} = \left[\overline{d_1}, \overline{d_2} \right]^T \quad \text{.....(3.7)}$$

que maximiza $c(d)$ sobre $R(p_l)$:

$$\overline{d} = \arg\max_{d \in \Re} \{ c(d) \} \quad \text{.....(3.8)}$$

Como salida se obtiene un vector de disparidades (el mapa de disparidad), uno para cada pixel de I_l.

3.1.4 Programación dinámica

La técnica de la programación dinámica (figura 3.4) fue propuesta por Cox en [CHR96], dicha técnica se utiliza para determinar si hay posibilidad de modificar las decisiones durante cierto periodo. También, se ocupa de los problemas en los que el tiempo no es una variable significativa. Un ejemplo de esta clase de problemas es cuando se debe tomar una decisión que requiera la distribución de una cantidad fija de recursos entre cierto número de usos alternativos. Ese tipo de problemas pueden resolverse descomponiéndolo en varias etapas, y de ese modo la decisión final se maneja como si fuera una serie de decisiones dependientes en el transcurso del tiempo.

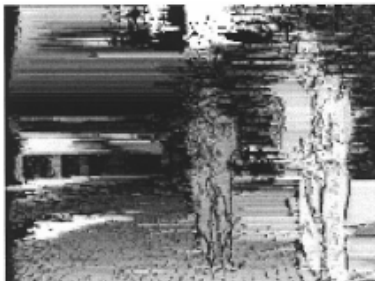


Figura 3.4.-Mapa de disparidad obtenido tras aplicar el algoritmo de programación dinámica [SCB03].

Aunque ese tipo de problema no se ocupa del factor tiempo por sí mismo, va ligado; sin embargo, la característica fundamental de la programación dinámica es la toma de decisiones en múltiples etapas. La idea que subyace en la programación dinámica es bastante simple: evitar calcular dos veces una misma operación, normalmente manteniendo una tabla de resultados conocidos que se llena a medida que se resuelven los subcasos. Esta es una técnica ascendente, basada en la división para encontrar un bloque similar. Normalmente se empieza por los subcasos más pequeños y por tanto más sencillos, combinando sus soluciones, se obtienen las respuestas para subcasos de tamaños cada vez mayores, hasta que finalmente se llega a la solución del caso original. En los problemas de visión estéreo para satisfacer la restricción de orden y no caer en el problema de multicorrespondencias, se puede utilizar programación dinámica para resolver las distintas correspondencias entre las imágenes, al mismo tiempo que proporciona la información correspondiente a oclusiones y discontinuidades.

Mediante una matriz se representa el camino mínimo que iguala los puntos de la imagen izquierda con los de la imagen derecha, pertenecientes a la misma línea epipolar (figura 3.5). Cada uno de los puntos de la imagen izquierda estaría representado por las filas y los puntos de la imagen derecha por las columnas. La idea es encontrar, para cada uno de los píxeles de la imagen izquierda, el píxel más similar de la imagen derecha [CSR03].

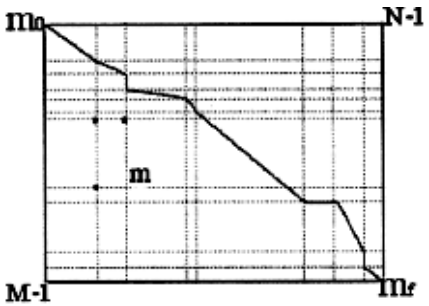


Figura 3.5.- Igualación estéreo utilizando programación dinámica [SCB03].

Se denotan a p_l y p_r como los píxeles de las imágenes izquierda y derecha respectivamente, además como $p_l = [i_l, j_l]$ y $p_r = [i_r, j_r]$ a sus coordenadas correspondientes y a $I_l = (i_l, j_l)$ y $I_r = (i_r, j_r)$ como los valores de intensidad respectivos. Cuando se utiliza la programación dinámica, se generan tantas matrices como líneas epipolares contenga la imagen, donde cada matriz almacena el resultado de las correspondencias de la línea epipolar de la imagen izquierda con su correspondiente línea epipolar de la imagen derecha. En cada matriz se consigue el camino mínimo entre los píxeles de la misma línea epipolar, esto es, las mejores igualaciones. Debido a la restricción de orden no se producen ciclos, ni recorridos hacia atrás, únicamente se admiten correspondencias con el píxel que ocupa la misma posición o la siguiente.

Utilizando programación dinámica se mantienen las restricciones de:

- Unicidad. Un píxel de la imagen izquierda no puede emparejar con más de un píxel de la imagen derecha y viceversa.

- Orden. Si un píxel z_{l1} empareja con z_{r1} entonces el siguiente pixel z_{l1+1} sólo puede emparejar con z_{r1+i} para cualquier $i > 0$.

El mínimo coste $C(m)$ de un camino desde *p* a *m* se define recursivamente según la ecuación 3.9.

$$C(m) = \min_{V_m}\left[c(p,m) + C(p)\right] \qquad \text{.....(3.9)}$$

Donde V_m es el conjunto de vecinos de *m*.

3.2 Algoritmo propuesto

3.2.1 Antecedentes

Ronald Aymert Fisher introdujo el análisis de la varianza (ANOVA), técnica para probar el significado de los datos procedentes de los experimentos y en un primer momento se interesó en experimentos por bloques aleatorios [Mon91]. El método consiste en separar matemáticamente el efecto y el error, si el experimento revelaba un efecto real, el método mostraría la fuerza de ese efecto en relación con el error. Dicho método se basa en los siguientes tres principios básicos:

1. Aleatorizar. Todos aquellos factores no controlados en el diseño experimental y que pueden influir en los resultados serán asignados al azar a las unidades experimentales.

2. Bloquear. Dividir o particionar las unidades experimentales en grupos llamados bloques de tal modo que las observaciones realizadas en cada bloque se realicen bajo condiciones experimentales lo más parecidas posibles. A diferencia de lo que ocurre con los tratamientos, el experimentador no está interesado en investigar las posibles diferencias de la respuesta entre los niveles del bloque.

3. Diseño Factorial. estrategia experimental que consiste en cruzar los niveles de todos los factores de tratamiento en todas las combinaciones posibles.

Cuando se desea un diseño en bloques o con factor bloque, se deben agrupar las unidades experimentales en bloques, a continuación se determina la distribución de los tratamientos en cada bloque y, por último, se asigna al azar las unidades experimentales a los tratamientos dentro de cada bloque. El modelo matemático es:

Respuesta = Constante + Efecto Tratamiento + Efecto Bloque + Error

Se define como:

- Factor a las variables independientes relacionadas con una variable de respuesta
- Nivel como el valor que un factor asume en un experimento (grado de intensidad).
- Tratamiento a la combinación específica de niveles de los factores que intervienen en un experimento
- Bloque al grupo de *i* unidades experimentales tan parecidas como sea posible con respecto a la variable *j*, asignándose aleatoriamente cada tratamiento a una unidad dentro de cada bloque.

Las ventajas que tiene este tipo de análisis son las siguientes:

- Que el análisis estadístico es relativamente simple.
- Un bloque puede incrementar la precisión, removiendo una fuente de variación proveniente del error experimental.
- Cualquier bloque será usado tantas veces como tratamientos se hayan agregado al diseño.

Aunque posee ciertas desventajas como son:

- El perder un dato puede causar alguna dificultad en el análisis.
- La asignación de tratamientos por error a unidades en un bloque incorrecto puede ocasionar problemas en el mismo.
- El diseño es menos eficiente que otros, en presencia de más de una fuente de variación; y que la eficiencia en el diseño decrece con el número de tratamientos y se incrementa con el número de bloques.

Con todo lo anteriormente mencionado este tipo de análisis provee estimaciones imparciales de las medias para las categorías de bloques, proporcionando información adicional del experimento, además de proporcionar una precisión satisfactoria en la mayoría de los casos sin el uso de un diseño más complejo.

3.2.2 Bases estadísticas

El modelo estadístico de este diseño es para un sistema de *a* cámaras que obtienen *a* imágenes en *b* bloques de pixeles:

$$I_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \begin{cases} i = 1, 2, ..., a \\ j = 1, 2, ..., b \end{cases}$$

En donde μ es una media general, τ_i es el efecto del de la i-ésima imagen, β_j es el efecto del j-ésimo pixel y ε_{ij} es el término usual $\eta\left(0, \sigma^2\right)$ de error aleatorio. Por lo que I_{ij} es la intensidad del j-ésimo pixel de la i-ésima imagen. Inicialmente se considera que tanto las imágenes como los pixeles son factores fijos. Más aún los efectos de las imágenes y de los pixeles se consideran como desviaciones de la media general, por lo tanto:

$$\sum_{i=1}^{a} \tau_i = 0 \qquad \sum_{j=1}^{b} \beta_j = 0$$

Lo que se desea es probar la igualdad de las medias de las intensidades de los pixeles en las imágenes. Así la hipótesis nula general es:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_a$$
$$H_1 : \text{al menos una } \mu_i \neq \mu_j$$

Una forma equivalente de expresar la hipótesis anterior, es la ecuación 3.10, donde se expresa la media del i-ésima imagen.

$$\mu_i = \frac{1}{b} \sum_{i=1}^{b} \mu + \tau_i + \beta_j = \mu + \tau_i \qquad \text{.....(3.10)}$$

Ahora si se expresa la misma ecuación 3.10, pero ahora en términos de los efectos sobre la Intensidad, es decir, de la i-ésima imagen y del j-ésimo quedaría como:

$$H_0 : \tau_1 = \tau_2 = ... = \tau_a = 0 \quad H_0 : \beta_1 = \beta_2 = ... = \beta_b = 0$$
$$H_1 : \tau_i \neq 0 \text{ al menos una } i \quad H_1 : \beta_j \neq 0 \text{ al menos una } j$$

Sean:

- $I_{i.}$ el total de las observaciones de la intensidad de la imagen i (Ecuación 3.11).
- $I_{.j}$ el total de las observaciones de la intensidad del pixel j (Ecuación 3.12).
- $I_{..}$ es el total de todas las observaciones en la intensidad o también $N=ab$ el número total de observaciones (Ecuación 3.13).

$$I_{i.} = \sum_{j=1}^{b} I_{ij} \qquad \text{i} = 1,2,.....,\text{a} \qquad \text{……...……...……..(3.11)}$$

$$I_{.j} = \sum_{i=1}^{a} I_{ij} \qquad \text{j} = 1,2,.....,\text{b} \qquad \text{……...……...……..(3.12)}$$

$$I_{..} = \sum_{i=1}^{a}\sum_{j=1}^{b} I_{ij} = \sum_{i=1}^{a} I_{i.} = \sum_{j=1}^{b} I_{.j} \qquad \text{……...……...……..(3.13)}$$

Similarmente:

- $\bar{I}_{i.}$ es el promedio de las Intensidades de la imagen i (Ecuación 3.14).
- $\bar{I}_{.j}$ es el promedio de las intensidades del pixel j (Ecuación 3.15).
- $\bar{I}_{..}$ es el promedio de todas las Intensidades (Ecuación 3.16).

$$\bar{I}_{i.} = \frac{I_{i.}}{b} \qquad \text{……...……...……..(3.14)}$$

$$\bar{I}_{.j} = \frac{I_{.j}}{a} \qquad \text{……...……...……..(3.15)}$$

$$\bar{I}_{..} = \frac{I_{..}}{N} \qquad \text{……...……...……..(3.16)}$$

La suma total de cuadrados corregida puede expresarse como en la ecuación 3.17.

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\left(I_{ij} - \bar{I}_{..}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\left[\left(I_{i.} - \bar{I}_{..}\right) + \left(I_{.j} - \bar{I}_{..}\right) + \left(I_{ij} - \bar{I}_{i.} - \bar{I}_{.j} + \bar{I}_{..}\right)\right]^2 \quad \text{...(3.17)}$$

Después de algunos pasos algebraicos simples pero tediosos, se comprueba que los tres términos que contienen productos cruzados son iguales a cero. Por lo tanto se deduce la ecuación 3.18. La ecuación 3.19 representa una suma total de cuadrados. Expresando simbólicamente las sumas de cuadrados (SS) de la ecuación 3.18.

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\left(I_{ij}-\bar{I}_{..}\right)^2 = b\sum_{i=1}^{a}\left(I_{i.}-\bar{I}_{..}\right)^2 + a\sum_{j=1}^{b}\left(I_{.j}-\bar{I}_{..}\right)^2 + \\ + \sum_{i=1}^{a}\sum_{j=1}^{b}\left(I_{ij}-\bar{I}_{i.}-\bar{I}_{.j}+\bar{I}_{..}\right)^2 \qquad \text{……...(3.18)}$$

$$SS_T = SS_{\text{Imágenes}} + SS_{Pixeles} + SS_E \qquad \text{……...(3.19)}$$

Ya que existen N observaciones de intensidad la SS_T tiene N-1 grados de libertad. La $SS_{\text{Imágnes}}$ y la $SS_{Pixeles}$ tienen a-1 y b-1 grados de libertad, respectivamente, porque existen *a* imágenes y *b* bloques pixeles. La suma de los cuadrados del error no es más que la suma de cuadrados total, menos la suma de cuadrados de las imágenes y de la de los pixeles. Existen *ab* celdas con *ab-1* grados de libertad entre ellas, por lo tanto, SS_E tiene *ab-1-(a-1)-(b-1)* grados de libertad. Considerando que las imágenes y los pixeles son fijos, puede mostrarse en las ecuaciones 3.20, 3.21 y 3.22 cuales son los valores esperados de las medias de los cuadrados.

$$E(MS_{\text{Imágenes}}) = \sigma^2 + \frac{b\sum_{i=1}^{a}\tau_i^2}{a-1} \qquad \text{……...(3.20)}$$

$$E(MS_{Pixeles}) = \sigma^2 + \frac{a\sum_{j=1}^{b}\beta_j^2}{b-1} \qquad \text{……...(3.21)}$$

$$E(MS_E) = \sigma^2 \qquad \text{……...(3.22)}$$

Por lo tanto, para probar la igualdad en las medias de imágenes y de pixeles, hay que usar los estadígrafos de la ecuación 3.23 y 3.24, respectivamente.

$$F_0 = \frac{MS_{\text{Imágenes}}}{MS_E} \qquad \text{……...(3.23)}$$

$$F_0 = \frac{MS_{Pixeles}}{MS_E} \qquad \text{……...(3.24)}$$

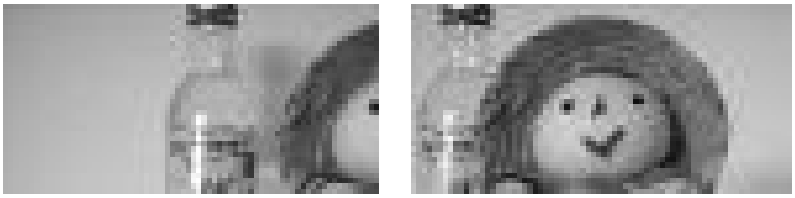
La Hipótesis nula es la verdadera, si la ecuación 3.23 tiene una distribución de probabilidad F de Fisher menor a $F_{\alpha,\kappa,\lambda}$, donde α es el intervalo de confianza, κ son los grados de libertad de $SS_{\text{Imágnes}}$ o $SS_{Pixeles}$ y λ son los grados de libertad de SS_E. La región crítica es el extremo superior de la Distribución F y se debería rechazar H_0, en el caso de las imágenes, si $F_0 > F_{\alpha,a-1,(a-1)(b-1)}$. También puede ser de interés la comparación entre las medias de los pixeles, por si existe una gran diferencia entre ellos. Al analizar los valores esperados de las medias de cuadrados puede parecer que la hipótesis $H_0 : \beta_i = 0$ puede probarse comprobando el estadígrafo expresado en la ecuación 3.24 con $F_{\alpha,b-1,(a-1)(b-1)}$. La

tabla 3.1 resume el estudio anteriormente desarrollado, lo que es conocido como una tabla de análisis de la varianza o tabla ANOVA.

FUENTE DE VARIACIÓN	**SUMA DE CUADRADOS**	**GRADOS DE LIBERTAD**	**MEDIA DE CUADRADOS**	F_0
Imágenes	$\sum_{i=1}^{a} \frac{I_{i.}^2}{b} - \frac{I_{..}^2}{N}$	$a-1$	$\frac{SS_{\text{Imágenes}}}{a-1}$	$\frac{MS_{\text{Imágenes}}}{MS_E}$
Pixeles	$\sum_{j=1}^{b} \frac{I_{.j}^2}{a} - \frac{I_{..}^2}{N}$	$b-1$	$\frac{SS_{Pixeles}}{b-1}$	$\frac{MS_{Pixeles}}{MS_E}$
Error	$SS_T - SS_{\text{Imágenes}} - SS_{Pixeles}$	$(a-1)(b-1)$	$\frac{SS_E}{(a-1)(b-1)}$	
Total	$\sum_{i=1}^{a}\sum_{j=1}^{b} I_{ij}^2 - \frac{I_{..}^2}{N}$	$N-1=ab-1$		

Tabla 3.1.- Análisis de la varianza para *a* Imágenes y *b* Pixeles.

3.2.3 Desarrollo



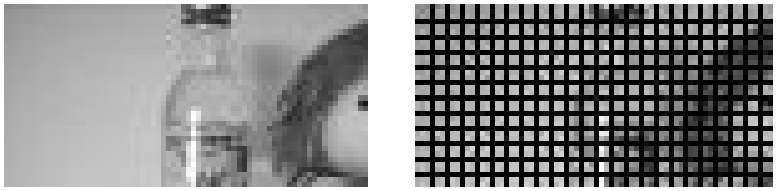
(a) (b)

Figura 3.6.- (a) Imagen Izquierda original 72 x 36 pixeles.
(b) Imagen Derecha original 72 x 36 pixeles

El algoritmo se basa en calcular la suma de cuadrados de las imágenes, pixeles y la interacción entre estas últimas o también llamado el error, las cuales indican seis propiedades que tienen los bloques de pixeles no traslapados de las imágenes, asumiendo un modelo paralelo en las cámaras, por lo que la búsqueda de igualaciones se realizará únicamente en forma horizontal. El par estéreo que se utilizará en lo sucesivo para desarrollar los ejemplos es el mostrado en la figura 3.6. Este algoritmo consta de 5 pasos los cuales se enuncian a continuación:

Bloqueo:

La imagen izquierda de $X \times Y$ pixeles (Figura 3.7a), es dividida en $m \times n$ bloques, por lo que cada bloque es de $X/m \times Y/n$ pixeles (Figura 3.7b).



(a) (b)

Figura 3.7.- (a) Imagen Izquierda original 72 x 36 pixeles.
(b) Imagen Izquierda dividida en 24 x12 bloques de 3 x 3 pixeles

Indexación:

La idea de este paso es dar un número a cada bloque, cuya fórmula general se encuentra expresada en la ecuación 3.25.

$$Bloque = \frac{n}{Y}x + \frac{m^2}{X}y \qquad \text{……...……...……..(3.25)}$$

Donde:

x es la posición horizontal del pixel objetivo (superior izquierdo dentro de un bloque, rojo). 
y es la posición vertical del pixel objetivo.

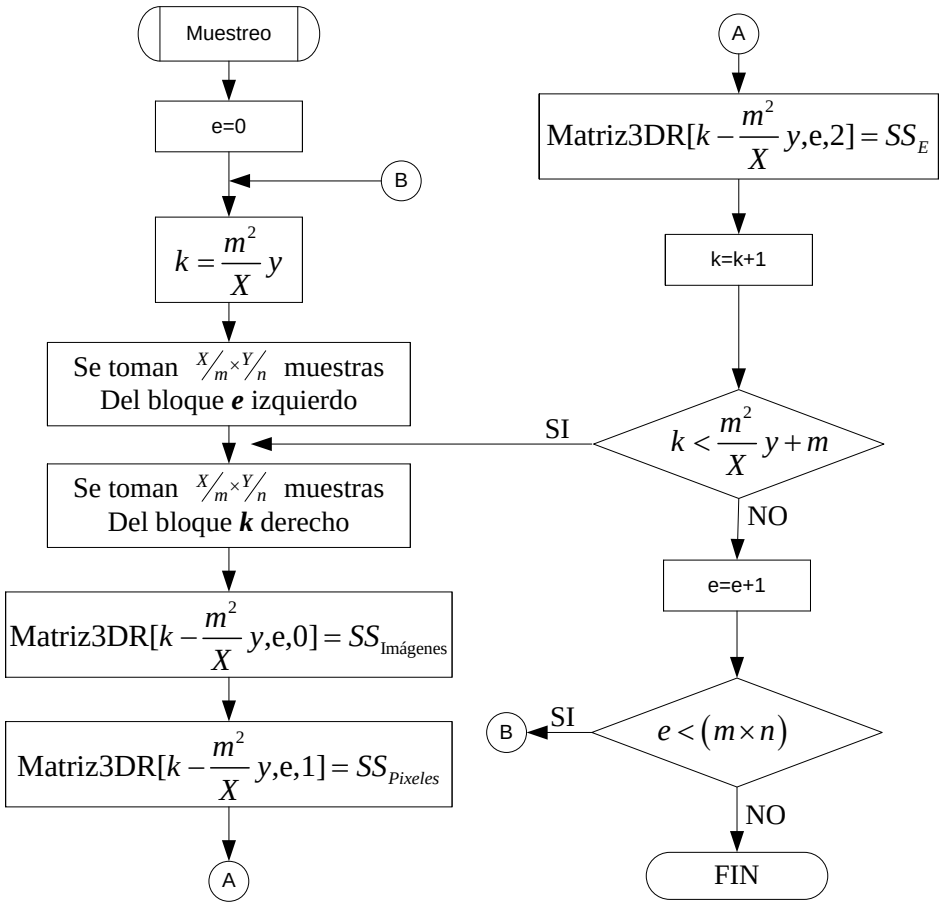
Por ejemplo, en la Figura 3.9a existe un bloque señalado en verde, este tiene las coordenadas del pixel objetivo en (69,18) y utilizando la ecuación 3.25 se obtiene lo siguiente:

$$Bloque = \frac{12}{36}(69) + \frac{24^2}{72}(18) = 167$$

Entonces dicho pixel objetivo está, junto con otros 8 pixeles, en el bloque 167 de la imagen izquierda.

Toma de muestras:

Una vez que se ha dividido la imagen y se sabe qué número posee cada bloque de la imagen izquierda, se procede a tomar las muestras para procesar el modelo estadístico propuesto, almacenándolas en Matriz3DR (Matriz tridimensional, Figura 3.8). Y para ello se utiliza el siguiente diagrama de flujo:



Cabe mencionar que este diagrama se diseña sólo para obtener *m* muestras sobre la línea epipolar $\frac{m}{X}y$ derecha y que se deberán incrementar o decrementar las coordenadas (x, y) de los pixeles, según se requiriese.

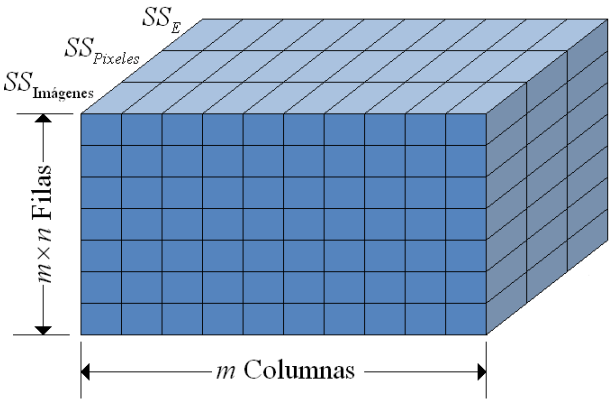


Figura 3.8.- Composición de Matriz3DR.

Ejemplo:

Paso 1: Después de varias iteraciones y que *e* tome el valor 167, por estar el pixel objetivo izquierdo en las coordenadas (69,18).

Paso 2: entonces *k* tomará el valor de 144 para tomar muestras sobre la línea epipolar 6 derecha.

Paso 3: se toman las 9 muestras del bloque *167* izquierdo.

Paso 4: se toman las 9 muestras del bloque 153 derecho, por ejemplo.

Paso 5: se calculan, mediante un análisis de varianzas, los estadígrafos $SS_{\text{Imágenes}}, SS_{Pixeles}$ & SS_E y se almacenan en Matriz3DR. Basándose en el experimento del Paso de Comparación, caso 2, donde $SS_{\text{Imágenes}} = 0, SS_{Pixeles} = 50516.4$ & $SS_E = 0$, se obtienen los valores que se almacenarán en la Matriz3DR, es decir, $\text{Matriz3DR}[9,167,0] = 0$, $\text{Matriz3DR}[9,167,1] = 50516.4$ y $\text{Matriz3DR}[9,167,2] = 0$.

Paso 6: se incrementa del valor del bloque derecho *k* entonces ahora *k*=154.

Paso 7: como *k* es menor que 168 bloques (muestras sobre la línea epipolar 6 derecha) se salta al paso 4 de caso contrario se continuaría.

Paso 8: cuando termine de rastrear toda la línea epipolar 6, el bloque izquierdo *e* se incrementa ahora *e*=168.

Paso 9: como *e* es menor que los 288 bloques totales, se saltará al paso 2 hasta que se rastreen todas las líneas epipolares, cuando esto suceda se terminará.

Comparación:

Cada uno de los tres estadígrafos tiene dos posibles valores, cero o que tiende a cero, con lo que se obtienen seis diferentes interpretaciones de dichos valores, las cuales se presentan en la tabla 3.2.

VALOR	INTERPRETACIÓN
$SS_{\text{Imágenes}} = 0$	Las intensidades de cada pixel en la imagen izquierda son iguales a las de la imagen derecha, aunque no necesariamente en la misma posición.
$SS_{\text{Imágenes}} \to 0$	Entre más cercano sea a cero, la intensidad de los pixeles es más parecida en los bloques izquierdo y derecho entre sí.
$SS_{Pixeles} = 0$	Las intensidades de pixeles izquierdos son las mismas, al igual que los derechos, aunque no necesariamente el mismo valor.
$SS_{Pixeles} \to 0$	Mientras más cercano a cero, más parecida es la intensidad de los bloques entre sí.
$SS_E = 0$	Los pixeles del bloque derecho e izquierdo, coinciden tanto en la posición como en la intensidad.
$SS_E \to 0$	Mientras más cercano a cero, más pixeles del bloque izquierdo coincidirán en el derecho.

Tabla 3.2.- Valores de estadígrafos empleados.

Con los anteriores valores se pueden construir hasta seis casos de estudio los cuales son:

Caso 1:

Cuando: $SS_{\text{Imágenes}} = 0$ $SS_{Pixeles} = 0$ $SS_E = 0$

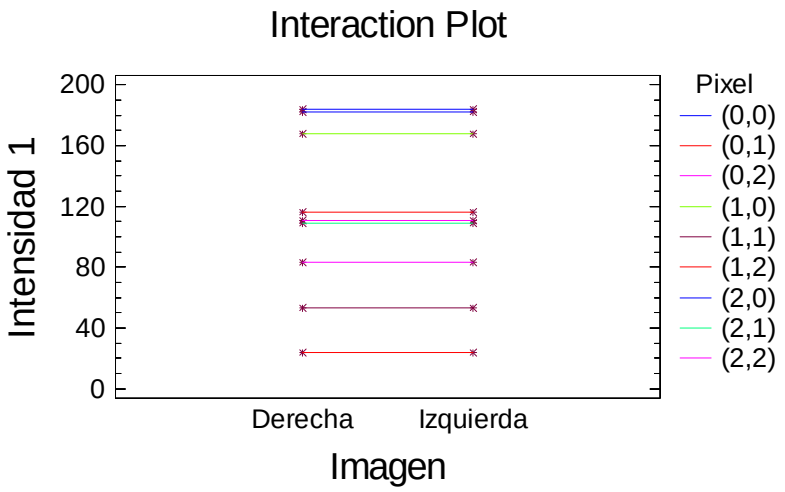
Esto significa que la intensidad de todos los pixeles de bloque izquierdo es la misma y que también es la aparece en el bloque izquierdo.

Imagen		Izquierda	Derecha
Pixel	(0,0)	184	184
	(0,1)	184	184
	(0,2)	184	184
	(1,0)	184	184
	(1,1)	184	184
	(1,2)	184	184
	(2,0)	184	184
	(2,1)	184	184
	(2,2)	184	184

Caso 2:

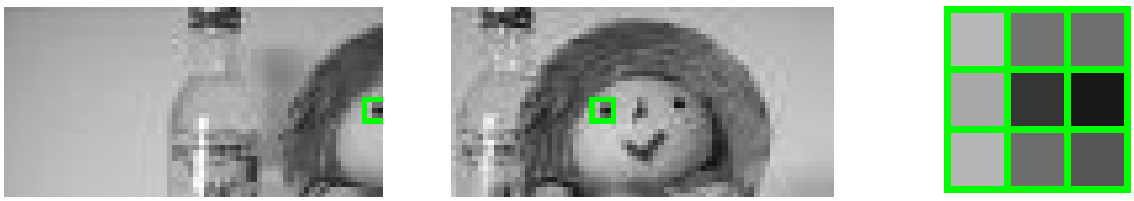
Cuando: $SS_{\text{Imágenes}} = 0$ $SS_{Pixeles} \to 0$ $SS_E = 0$

Imagen		Izquierda	Derecha
Pixel	(0,0)	184	184
	(0,1)	116	116
	(0,2)	111	111
	(1,0)	168	168
	(1,1)	53	53
	(1,2)	24	24
	(2,0)	182	182
	(2,1)	109	109
	(2,2)	83	83



Source	Sum of Squares
MAIN EFFECTS	
A:Imagen	0.0
B:Pixel	50516.4
INTERACTIONS	
AB (ERROR)	0.0
TOTAL (CORRECTED)	50516.4

Los datos presentados para este caso, son las intensidades reales de la figura 3.9 de los bloques 167 izquierdo y encontrado en el 153 derecho. Lo representado significa que cada pixel de la imagen izquierda corresponden con su respectivo par en la imagen derecha cuando la suma de cuadrados tanto de la imagen como del error sea igual a cero. Y mientras más cercano a cero sea la suma de cuadrados de los pixeles más parecidas son las intensidades.



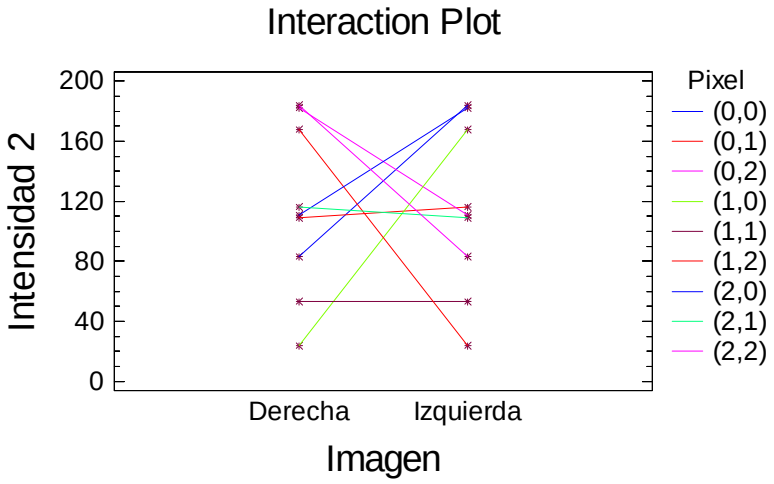
(a) (b) (c)

Figura 3.9.- (a) Imagen Izquierda original señalando el bloque 167 de 3 x 3. (b) Imagen derecha original señalando el bloque 153 de 3 x 3. (c) Bloque 167 izquierdo y 153 derecho ampliado.

Caso 3:

Cuando $SS_{\text{Imágenes}} = 0$ $SS_{Pixeles} \to 0$ $SS_E \to 0$

Imagen		Izquierda	Derecha
Pixel	(0,0)	184	83
	(0,1)	116	109
	(0,2)	111	182
	(1,0)	168	24
	(1,1)	53	53
	(1,2)	24	168
	(2,0)	182	111
	(2,1)	109	116
	(2,2)	83	184



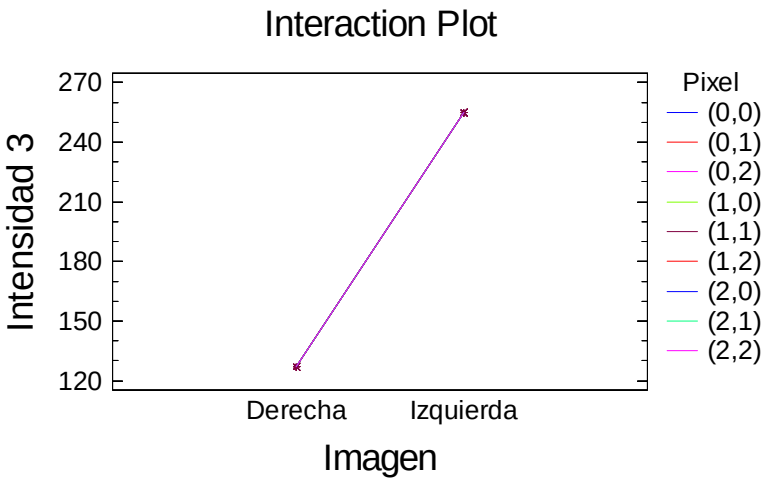
Source	Sum of Squares
MAIN EFFECTS	
A:Imagen	0.0
B:Pixel	14489.4
INTERACTIONS	
AB (ERROR)	36027.0
TOTAL (CORRECTED)	50516.4

Cuando la suma de cuadrados de la imagen sea igual a cero y la suma de cuadrados tanto del pixel como del error tiendan a cero, significa que existen las mismas intensidades en los pixeles de la izquierda en la derecha aunque no coincidan en su posición.

Caso 4:

Cuando $SS_{\text{Imágenes}} \to 0 \quad SS_{Pixeles} = 0 \quad SS_E = 0$

Imagen		Izquierda	Derecha
Pixel	(0,0)	255	127
	(0,1)	255	127
	(0,2)	255	127
	(1,0)	255	127
	(1,1)	255	127
	(1,2)	255	127
	(2,0)	255	127
	(2,1)	255	127
	(2,2)	255	127



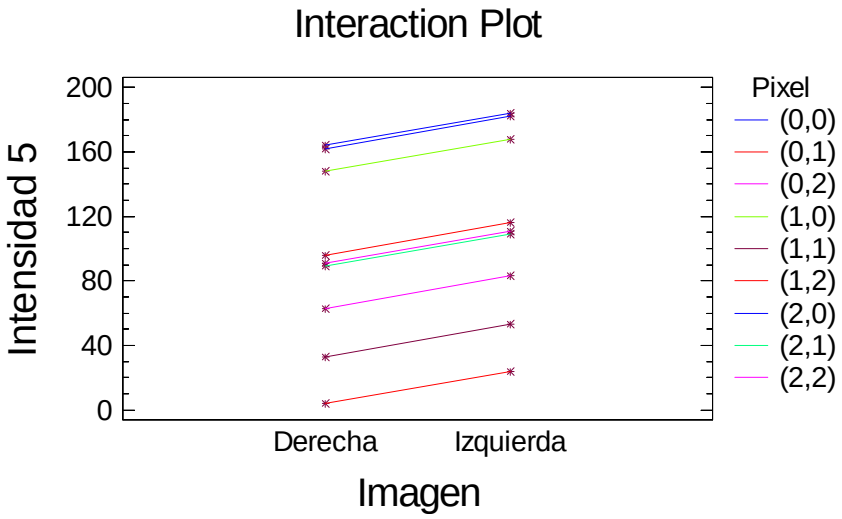
Source	Sum of Squares
MAIN EFFECTS	
A:Imagen	73728.0
B:Pixel	0.0
INTERACTIONS	
AB (ERROR)	0.0
TOTAL (CORRECTED)	73728.0

Cuando la suma de cuadrados de la imagen tienda a cero y la suma de cuadrados tanto del pixel como del error sean igual a cero, significa que existe la misma intensidad en todo el bloque de los pixeles izquierdos y de igual manera la intensidad de los pixeles derechos es la misma pero dichas intensidades son diferentes entre sí. Además entre más cercana a cero sea la suma de cuadrados de la imagen más parecidos son los pixeles.

Caso 5:

Cuando $SS_{\text{Imágenes}} \to 0 \quad SS_{Pixeles} \to 0 \quad SS_E = 0$

Imagen		Izquierda	Derecha
Pixel	(0,0)	184	164
	(0,1)	116	96
	(0,2)	111	91
	(1,0)	168	148
	(1,1)	53	33
	(1,2)	24	4
	(2,0)	182	162
	(2,1)	109	89
	(2,2)	83	63



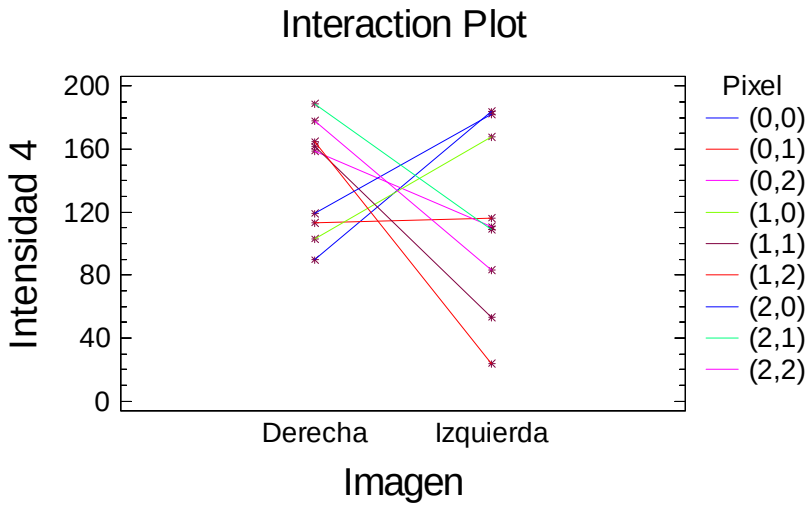
```
Source                 Sum of Squares
-------------------------------------
MAIN EFFECTS
 A:Imagen                     1800.00
 B:Pixel                      50516.4

INTERACTIONS
 AB (ERROR)                      0.00
 -------------------------------------
TOTAL (CORRECTED)             52316.4
```

Cuando la suma de cuadrados, tanto de la imagen como del pixel, tiendan a cero y la suma de cuadrados del error sea igual a cero, significa que el bloque izquierdo es igual al derecho, pero en dicho bloque derecho existe un cambio de iluminación, el cual es constante. En ejemplo mostrado se le restó 20 a cada pixel del bloque izquierdo para obtener el derecho.

Caso 6:

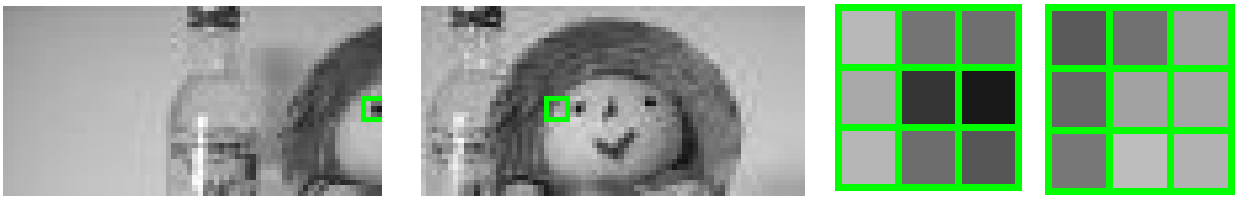
Cuando $SS_{\text{Imágenes}} \to 0 \quad SS_{Pixeles} \to 0 \quad SS_E \to 0$

Imagen		Izquierda	Derecha
Pixel	(0,0)	184	90
	(0,1)	116	113
	(0,2)	111	159
	(1,0)	168	103
	(1,1)	53	162
	(1,2)	24	165
	(2,0)	182	119
	(2,1)	109	189
	(2,2)	83	178




```
Source                 Sum of Squares
-------------------------------------
MAIN EFFECTS
 A:Imagen                     3416.89
 B:Pixel                      5728.11

INTERACTIONS
 AB (ERROR)                  29846.10
 -------------------------------------
TOTAL (CORRECTED)            38993.10
```

Los datos presentados para este caso, son las intensidades reales de la figura 3.10 de los bloques 167 izquierdo, que es comparado con el 152 derecho. Lo representado es la suma de cuadrados tanto de la imagen como del error y del pixel cuando esta es diferente de cero o tenderá a cero, lo cual significa que los bloques son completamente diferentes.



(a) (b) (c) (d)

Figura 3.10.- (a) Imagen Izquierda original señalando el bloque 167 de 3x3. (b) Imagen derecha original señalando el bloque 152 de 3x3. (c) Bloque 167 izquierdo ampliado. (d) Bloque152 derecho ampliado.

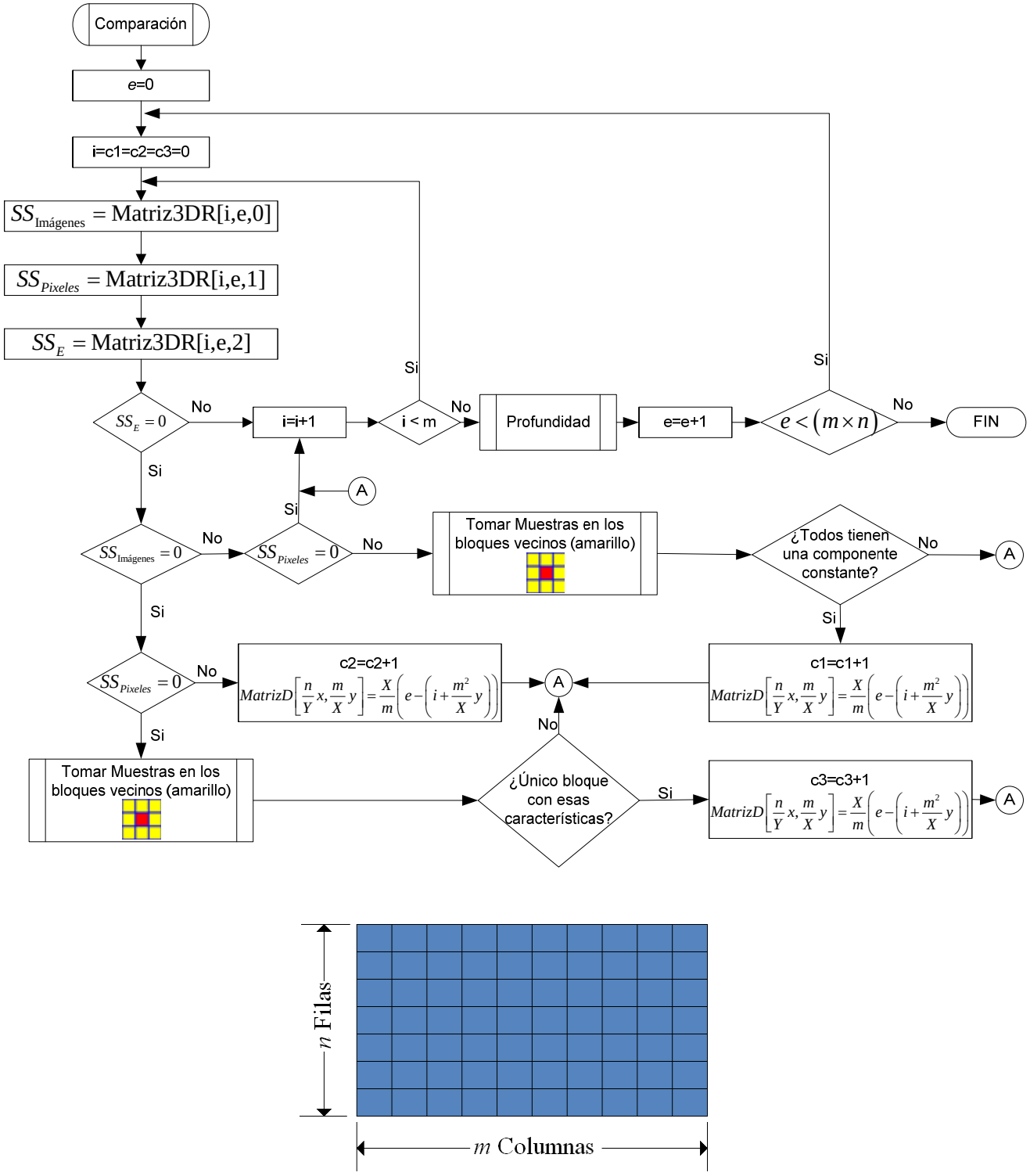


Figura 3.11.- Composición de MatrizD.

En el paso de comparación tiene como objetivo principal encontrar un cero o la menor cantidad en la sumatoria de cuadrados del error, lo que significa que las intensidades de los pixeles coinciden tanto en magnitud como en posición, aunque pueden tener una componente adicional por un cambio de luminosidad. En teoría sólo se debería de tener un solo bloque *e* izquierdo que corresponda en algún derecho. Cuando se encuentre el par estéreo de dicho bloque *e*, se calculará con la ecuación 3.26 su disparidad y se almacenará en una matriz llamada MatrizD (Figura 3.11), pero podría ocurrir que coincidiera más de un par estéreo sobre la línea epipolar rastreada, es por ello que se crearon los contadores c1, c2 y c3. La variable índice *i* es la componente en el eje x de la Matriz3DR, la cual se incrementa hasta *m* para comparar al bloque *e* por toda la línea epipolar $\frac{m}{X}y$ derecha.

Estimación de profundidad:

La profundidad entre bloques izquierdo y derecho pares, se calculará con la ecuación 3.26.

$$d = {}^{X}\!/_{m}\left(\#\,BloqueIzquierdo - \#\,Bloquederecho\right) \qquad \text{……...……...……..(3.26)}$$

Sustituyendo la ecuación 3.26 en la ecuación 2.1 y utilizando el ejemplo presentado en la figura 3.7, donde *s= 0.06 m y f=140 pixeles.*

$$z = \frac{(0.06)\times(140)}{\frac{72}{24}(167-153)} = \frac{8.4}{3\times 14} = \frac{8.4}{42} = 0.2$$

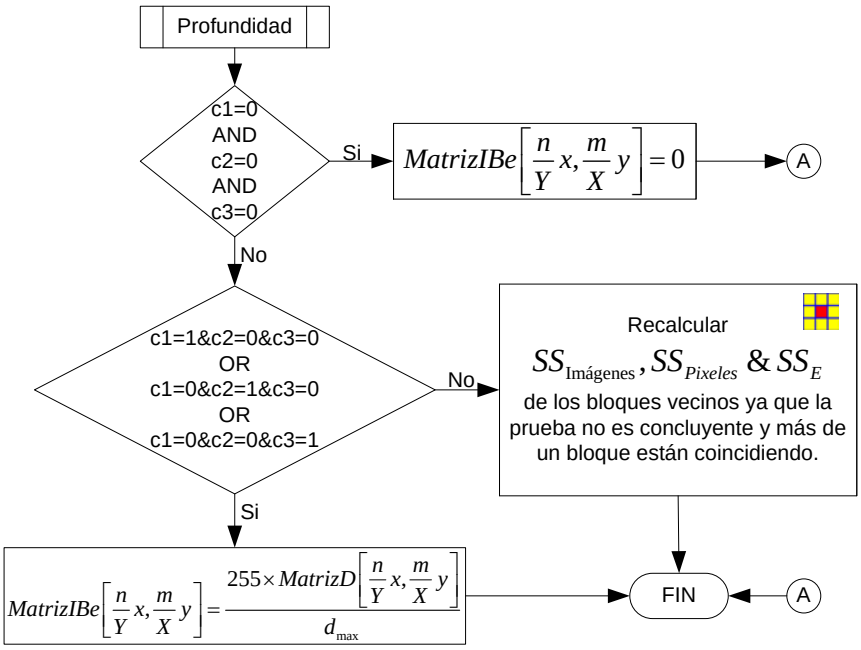
Por lo tanto el punto donde se encuentran los bloques izquierdo 167 y derecho 153 está a 20 centímetros de la cámara. Con todas las disparidades de todos los bloques pares e impares (a los cuales se les asignará la máxima profundidad) se puede construir un mapa de profundidad utilizando la ecuación 3.27. Tomando en cuenta que el valor de máxima distancia es un cero o negro y el de mínima distancia un dos cientos cincuenta y cinco o blanco.

$$IBe = \frac{255\times d_{be}}{d_{\max}} \qquad \text{……...……...……..(3.27)}$$

Donde: IBe , es la intensidad del bloque *e* izquierdo dentro del mapa de profundidad.

d_{be} , es la profundidad del bloque *e* izquierdo con respecto a su par derecho.

$d_{\max}$, es la profundidad máxima encontrada.



En la rutina de *Profundidad* primero se verifica si todos los contadores c1, c2, y c3 están en cero, lo que significaría que ningún bloque izquierdo correspondió en algún derecho. En el caso donde existiese algún contador diferente de cero, se verifica que sea único, es decir, que tenga valor uno y que los otros dos tengan un valor de cero. Si esto último ocurriese se calcula la intensidad correspondiente a la disparidad encontrada (Ecuación 3.27) y se almacenaría en *MatrizIBe*, pero si no indicaría que hubo algún contador con más de uno, dos contadores con uno o cualquier otro valor o los tres contadores con valores mayores o iguales que uno, por lo que hay que recalcular $SS_{\text{Imágenes}}, SS_{Pixeles}$ & SS_E pero ahora incluyendo a los bloques vecinos y obtener así una mejor conclusión.

Capítulo 4

Conclusiones

En resumen a lo largo de la presente memoria:

- Se conocieron los procedimientos de captura, representación computacional, codificación y visualización de imágenes naturales estereoscópicas.
- Se expusieron técnicas de estimación de profundidad.
- Se propuso una técnica diferente a las actuales que estime la profundidad de una escena estereoscópica.
- Se desarrolló un experimento que sustenta la teoría del algoritmo propuesto.

También se conocieron las fases y subfases de obtención de imágenes tridimensionales, después se expusieron técnicas para estimar la profundidad de un par de imágenes estéreo. A partir de los conocimientos adquiridos en el estudio de las técnicas de estimación de profundidad existentes, se propuso una técnica diferente la cual es escalable para el trabajo con colores.

En el segundo capítulo se explica el proceso de obtención de imágenes naturales tridimensionales, el cual como primera sub-fase tiene que capturar a las imágenes, ubicando objetos y cámaras utilizando la geometría epipolar, ya sea colocando dichas cámaras en un modelo paralelo o uno convergente. Se realizó una pequeña introducción de lo que significa estimar la profundidad, ya sea basada en regiones o haciendo uso de las propias características de las imágenes. También se expusieron tres técnicas de codificación y representación en tres dimensiones, la primera se basa en encontrar patrones recurrentes multiescalares, mientras que la segunda y tercera se fundamentan en la Transformada Wavelet, con el fin de que el lector perciba la necesidad de tener un adecuado algoritmo de codificación para no perder datos de profundidad. Aunque en la vida real estas primeras fases son invisibles para el usuario final, ya que por lo general las personas solamente se limitan a observar las imágenes y esperar que los objetos *salten* sobre ellas, por ello fue necesario exponer también las diferentes formas de visualización que existen, la automática, donde el espectador se coloca en un punto y las imágenes se orientan al ojo adecuado, y las que se ayudan de gafas polarizadas, anáglifo u obturadoras.

En el tercer capítulo se presentan cuatro algoritmos diferentes que estiman la profundidad. Algunos usaron raíces cuadradas, al calcular las desviaciones estándar o varianzas, esto amplia los ciclos de máquina para el cómputo de los algoritmos. El algoritmo propuesto, basado en un análisis de varianzas de bloques aleatorizados, sólo utiliza operaciones simples, como sumas, restas, divisiones y multiplicaciones, que al ser cíclicas se podrán implementar en lazos for, por citar un ejemplo. También se analizaron los casos generales en los que dos bloques de pixeles pudieron caer y así concluir si dichos bloques son idénticos, muy parecidos, con los mismos pixeles pero en posiciones diferentes, si eran homogéneos o inclusive si se requiriera si eran heterogéneos. No se contrastó el algoritmo presentado con ningún otro, en cuanto a los tiempos de ejecución, ya que la idea central de este trabajo de investigación es presentar un algoritmo

alternativo que facilite el cálculo de la disparidad y sea línea de exploración futura el desarrollo del software necesario.

Posibles líneas de investigaciones futuras para el tratamiento y codificación de escenas tridimensionales obtenidas a partir de pares de imágenes estereoscópicas son:

1. Implementación del software que calcule el mapa de profundidad utilizando el presente algoritmo, fijándose el objetivo principal de mejorar los tiempos de ejecución con respecto a otros algoritmos.

2. Utilizar características que arroja el algoritmo por si mismo, que en esta memoria no se utilizan, como son:

    a) Principio de Heterogeneidad, es decir, que se puede encontrar un bloque grande que contenga una serie de texturas diferentes, las cuales se podrán identificar.

    b) Bloqueo progresivo, se puede bloquear una zonas de acuerdo a las propiedades de la misma, pudiendo trabajar al mismo tiempo con bloques muy grandes y con bloques de unos cuantos pixeles.

3. Utilización de un sistema de compresión diferente que se adecue las características del algoritmo propuesto, este puede enfocarse en la teoría de fractales utilizando transformada wavelet, y así pasar de los detalles de mayor energía los de menor energía.

4. Implementar un código de corrección de errores [Wki01], para la transmisión de las imágenes tridimensionales naturales, el cual podría ser un código LDPC (Low-Density Parity-Check) embedded aplicado en las tramas de bits de salida. Esto es porque la mayoría de los algoritmos de codificación de imágenes estéreo carecen de una protección contra los errores en el canal de transmisión [TPM03, HAH07].

5. Diseñar el algoritmo con una análisis de covarianzas, para verificar cuales son los posibles casos de estudio.

6. Como se mencionó en el subtema 3.2.1, si en un modelo de análisis de varianzas de bloques aleatorizados se removiesen los factores que provoquen un fuerte variación, lo que se conseguiría es incrementar la precisión, por lo que se pueden implementar sistemas estereoscópicos convergentes o paralelos con una cantidad de *a* cámaras, que solamente procesen los bloques de imágenes relevantes, descartando los que aumenten el error significativamente.

Referencias

[BS00] N. Boulgouris & M. Strintzis. *Embedded coding of stereo images.* Extraído de Image Processing International Conference, volumen 3, páginas 640 - 643,Septiembre 2000.

[BS02] N. Boulgouris & M. Strintzis. *A family of wavelet-based stereo image coders.* Extraído de Circuits and Systems for Video Technology, IEEE Transactions, volumen 12, páginas 898 - 903, Octubre 2002.

[BT04] J. Blanchard & R. Tsuneto. *Stereoscopic viewing,* http://www.hitl.washington.edu/scivw/EVE/III.A.1.b.StereoscopicViewing.html, 2004.

[CCK07] J. Cho, I. Chang, & K. Seung-man. *Depth Image Processing Technique for Representing Human Actors in 3DTV using Single Depth Camera.* Extraído de 3DTV Conference, volumen 1, páginas 1 - 4, Mayo 2007.

[CG66] W. Campell & R.W. Gubisch. *Optical quality of human eye*. Extraído de Physiological Laboratory, University of Cambridge, páginas 558-578, 1966.

[CHR96] I. Cox, S. Hingorani & S. Rao. *A maximum likelihood stereo algorithm.* Extraído de Computer Vision and Image Understanding, volumen 63:3, páginas 542-567, 1996.

[CKC04] Choi, C.; Kwon, B. & Choi, M. A Real-Time Field-Sequential Stereoscopic Image Converter. Extraído de IEEE Transactions on Consumer Electronics, volumen 50, páginas 903-910, 2004.

[CSR03] P. Compañ, R. Satorre & R. Rizo. *Disparity estimation in stereoscopic vision by simulated annealing.* Extraído del Departament of Computer Science and Artificial Intelligence University of Alicante, volumen 1, páginas 1-8, 2003.

[CSS02] B. Chai, S. Sethuraman & H. Sawhney. *A depth map representation for real-time transmission and view-based rendering of a dynamic 3D scene.* Extraído de 3D Data Processing Visualization and Transmission, volumen 1, páginas 107 - 114, Junio 2002.

[CWR07] T. Chinapirom, U. Witkowski & U. Rückert. *Steroscopic Camera for Autonomous Mini-Robots Applied in KheperaSot League*, Extraído de Heinz Nixdorf Institute y University of Paderborn, volumen 1, páginas 1-6, 2007.

[DCS02] M. Duarte, M. Carvalho, E. da Silva, C. Pagliari & G. Mendonca. *Stereo image coding using multiscale recurrent patterns.* Extraído de Image Processing International Conference, volumen 2, páginas II-661 - II-664, Septiembre 2002.

[ESY08] L. Eun-Kyung, K. Sung-Yeol, & J. Young-Ki. *High-Resolution Depth Map Generation by Applying Stereo Matching Based on Initial Depth Informaton.* Extraído de 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, volumen 1, páginas 201 - 204, Mayo 2008.

[FCA07] A. Ferreira, R. Cruz, L. Antunes & D. Chadwick. *Access Control: how can it improve patients' healthcare?* Extraído de Studies in Health Technology and Informatics, volumen 127, páginas 127-139, Junio 2007.

[Fer05] J. Fernández. *El equipamiento para la fotografía digital*. Extrído de Rev Esp Ortod, páginas 75-84, 2005.

[GCA07] J. Gallego, P. Compañ, P. Arques, C. Villagrá, & R. Molina. *Detección de objetos y estimación de su profundidad mediante un algoritmo de estéreo basado en segmentación.* II Congreso Español de Informática,volumen 1, páginas 1-8, 2007.

[GHB08] L. Gustavo, K. Hari, & F. Borko. *3D Video Quality Evaluation with Depth Quality Variations.* Extraído de 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, volumen 1, páginas 301 - 304, Mayo 2008.

[Gri85] W. Grimson. *Computational Experiments with a Feature Based Stereo Algorithm*. Extraído de IEEE Transactions on Pattern Analysis and Machine Intelligence, volumen 7, páginas 17-34, 1985.

[HAH07] K. Hongt , Y. Ant & S. Hongt. *Remote Sensing Image CompressionBased on the Improvement of SPECK Algorithm.* Extraído de Proceedings of IWSDA, 2007.

[HZ04] R. Hartley & A. Zisserman. *Multiple View Geometry in Computer Vision.* Extraído de Cambridge University Press, páginas 344-360, 2004.

[IRP02] H. Isaa, Y. Ruichek & J. Postaire. *Extracting Depth Information from Stereo Linear Images using genetic Approach,* Extraído del First International IEE Symposium Inteligent Systems, volumen 8, páginas 185-189, 2002.

[JAP05] T. Jebara, A. Azarbayejani & A. Pentland. *3D structure.* Extraído de motion IEEE signal processing Magazine, 2005.

[JZS02] X. Jizheng, X. Zixiang & L. Shipeng. *High performance wavelet-based stereo image coding*. Extraído del Circuits and Systems IEEE International Symposium, volumen 2, páginas II-273 - II-276, Mayo 2002.

[LW97] J. Liang & D.R. Williams. *Aberrations and retinal image quality of the normal human eye.* Extraído de JOSA A, Vol. 14, Issue 11, páginas 2873-2883, 1997.

[Mig00] J. Migdal. *Depth Perception Using a Trinocular Camera Setup and Sub-Pixel Image-Correlation Algorithm*. Extraído del Mitsubishi Electric Research Laboratories, volumen 20, páginas 1-18 , Diciembre 2000.

[MK99] K. Mi-Hyun & S. Kwang-Hoon. *Edge-preserving directional regularization technique for disparity estimation of stereoscopic images*. Extraído de Consumer Electronics, IEEE Transactions, volumen 45, páginas 804 - 811, Agosto 1999.

[MMA05] J. J. Moreno, E. J. Martínez & E. Y. Aguilar. *Metodología para la Creación de Objetos de Aprendizaje de apoyo a la educación*. Extraído del 4° Congreso Internacional de Ingeniería Electromecánica y de Sistemas, volumen 4, páginas 1-6, Noviembre 2005.

[Mon91] D. Montgomery. *Diseño y Análisis de Experimentos.* Editado por Grupo Editorial Iberoamérica, Capítulos 3,9 y 11, 1991.

[Mor05] J. J. Moreno. *Metodología para la Creación de Objetos de Aprendizaje de apoyo a la educación*. Extraído de Memoria de tesis de Maestría, Instituto Politécnico Nacional de México, páginas 151-196, Agosto 2005.

[MP04] W. Matusik & H. Pfister. 3*D TV: A Scalable System for Real-Time Acquistion, Transmission and Autostereoscopic Display of Dynamic Scenes.* Extraído de ACM Transactions on Graphics SIGGRAPH,volumen 23, páginas 814-824, Agosto 2004.

[MPH79] D. Marr, T. Poggio & E. Hildreth. *The smallest channel in early human vision.* Extraído de Journal Optic American Soceity , volumen 70, páginas 868-870, 1979.

[NEB02] M. Nayan, E. Edirisinghe & H. Bez. *Baseline JPEG-like DWT CODEC for disparity compensated residual coding of stereo images.* Extraído de The 20th Eurographics UK Conference, volumen 1, páginas 67 - 74, Junio 2002.

[OA05] G. Overett & D. Austin. *Stereo Vision Motion Detection from a Moving Platform*. Extraído del Robotic Systems Lab Australia,volumen 1, páginas 1-11, 2005.

[Ols08] R. Olsson. *Empirical Rate-Distortion Analysis of JPEG 2000 3D and H. 264/AVC Coded Integral Imaging Based 3D-Images.* Extraído de 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, volumen 1, páginas 113 - 116, Mayo 2008.

[PL07] T. Pachidis & J. Lygouras. *Pseudostereo-Vision System: A Monocular Stereo-Vision System as a Sensor for Real-Time Robot Applications.* Extraído de Instrumentation and Measurement, IEEE Transactions, volumen 56, páginas 2547 - 2560, Diciembre 2007.

[PMF85] S. Pollard, J. Mayhew & J. Frisby. *Pmf: A stereo correspondence algorithm using a disparity gradiente limit.* Extraído de la revista Perception, volumen 14, páginas 449-420, 1985.

[RE06] F. Remondino & S. El-Hakim. *Image-Based 3d Modelling: A Review.*Extraído de Remote Sensing and Photogrammetry Society and Blackwell Publishing, volumen 21, páginas 269-291, 2006.

[SB05] S. Se, & M. Brady. *Stereo Vision-based Obstacle Detection for Partially Sighted People*. Extraído del Department of Engineering Science, University of Oxford volumen 1, páginas 1-8, 2005.

[SCB03] R. Satorre, P. Compañ & A. Botía. *Estimación de disparidad en visión estereoscópica mediante la integración de diversas técnicas combinadas con multirresolución.* Extraído de CAEPIA-TTIA 2003 / X Conferencia de la Asociación Española para la Inteligencia Artificial, V Jornadas de Transferencia Tecnológica de Inteligencia Artificial, volumen 27,páginas 1-4, 2003.

[SCM91] J.A. Silva, A. Campilho & J.M. dos Santos. *3-D data acquisition using the ratio of two intensity images.* Extraído de la 6th Mediterranean Electrotechnical Conference, volumen 2, páginas 1246 - 1250, Mayo 1991.

[SMT06] P. Schelkens, A. Munteanu, A. Tzannes, & C. Brislawn. *JPEG2000. Part 10. Volumetric data encoding*. Extraído de IEEE International Symposium onCircuits and Systems,volumen 2, páginas 3874-3877, Mayo 2006.

[SR03] R. Srikanth & A. Ramakrishnan. *Wavelet based coding of 2D and 3D MR images.* Extraído de TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, volumen 2, páginas 515 - 519,Octubre 2003.

[STS07] S. Sulaiman, N. Tahir, A. Shah & A. Hussain. *Human Motion Analysis using Virtual Reality*. Extraído de The 5th Student Conference on Research and Development, volumen 11, páginas 1-4, Diciembre 2007.

[THC05] J.R. Terán, S.M. Herrera & L. Corrales. *Prototipo de facomulsificador para el tratamiento de las cataratas.* Extraído de XIX Jornadas en Ingeniería Eléctrica y Electrónica, páginas 80-84, 2005.

[TPM03] X. Tang, W. A. Pearlman, & J. W. Modestino. *Hyperspectral Image Compression Using Three-Dimensional Wavelet Coding.* Extraído de SPIE/IS&T Electronic Imaging 2003, volumen 5022, 2003.

[TV98] E . Trucco & A. Verri. *Introductory Techniques for 3D Computer Vision.* Extraído de Prentice Hall, 1998.

[VMP04] A. Vetro, W. Matusik, H. Pfister & J. Xin. *Coding Approaches for End-to-End 3D TV Systems.* Extraído de Picture Coding Symposium, volumen 1, páginas 2-8, Diciembre 2004.

[WEB01] http://www.geocities.com/acarvajaltt/images/temas/ojo_humano.jpg

[WEB02] http://thales.cica.es/rd/Recursos/rd99/ed99-0273-01/dibujos/camara_oscura.jpg

[WEB03] http://www.freewebs.com/lindajudithramirez/000149770.png

[WEB04] http://manesweb.8k.com/

[WEB05] http://ciberhabitat.gob.mx/medios/camaras/funcionamiento.htm

[WEB06] http://ciberhabitat.gob.mx/medios/camaras/images/ccd.jpg

[WEB07] http://www.desarrolloweb.com/articulos/1865.php

[WEB08] http://www.desarrolloweb.com/articulos/images/diseno/10/

[WEB09] http://ciberhabitat.gob.mx/medios/camaras/images/fromatos.jpg

[WEB10] http://www.robots.ox.ac.uk/~vgg/hzbook/Fchapter8.html

[WEB11] http://www.planar3d.com

[WEB12] http://www.paralax.com.mx

[Whe38] C. Wheatstone. *Contributions to the Physiology of Vision. Part the First. On some remarkable, and hitherto unobserved, Phenomena of Binocular Vision,*1838.

[Wki01] http://en.wikipedia.org/wiki/Forward_error_correction

[WO00] W. Woontack & A. Ortega. *Overlapped block disparity compensation with adaptive windows for stereo image coding*. Extraído de Circuits and Systems for Video Technology, IEEE Transactions, volumen 10, páginas 194 - 200, Marzo 2000.

[XXL02] J. Xu, Z. Xiong & S. Li. *High performance wavelet-based stereo image coding.* Extraído de la Académia China de la Ciencia, volumen1, páginas 1-4, 2002.

[XZ96] G. Xu & Z. Zhang. *Epipolar Geometry in Stereo, Motion and Object Recognition.* Extraído de Kluwer Academic Publishers, volumen 1, páginas 1-10, 1996.

Anexo A

El Sistema Visual Humano y sus Analogías

A.1 Características del Sistema Visual Humano

El 50 % de la información que se recibe del entorno es a través de los ojos. La enorme cantidad de información que se recibe en un simple vistazo del entorno se guarda durante un segundo en la memoria y luego se desecha casi toda. El ojo humano es un sistema óptico formado por una dioptría esférica y una lente, que reciben, respectivamente, el nombre de córnea y cristalino, que son capaces de formar una imagen de los objetos sobre la superficie interna del ojo, en una zona denominada retina, que es sensible a la luz.

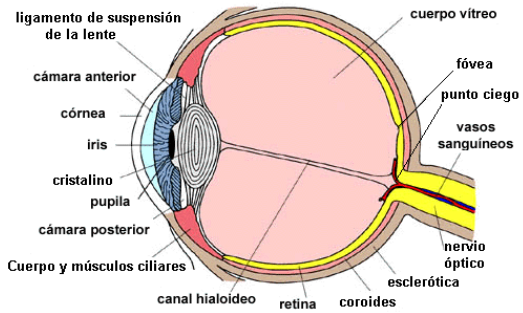


Figura A.1.- Partes principales del ojo [WEB01].

En la figura A.1 se ven claramente las partes que forman el ojo. Tiene forma aproximadamente esférica y está rodeado por una membrana llamada esclerótica que por la parte anterior se hace transparente para formar la córnea. Tras la córnea hay un diafragma, el iris, que posee una abertura, la pupila, por la que pasa la luz hacia el interior del ojo. El iris es el que define el color de los ojos y el que controla automáticamente el diámetro de la pupila para regular la intensidad luminosa que recibe el ojo. El cristalino está unido por ligamentos al músculo ciliar. De esta manera el ojo queda dividido en dos partes: la posterior que contiene humor vítreo y la anterior que contiene humor acuoso. El índice de refracción del cristalino es 1,437 y los del humor acuoso y humor vítreo son similares al del agua [THC05].

El cristalino enfoca las imágenes sobre la envoltura interna del ojo, la retina. Esta envoltura contiene fibras nerviosas (prolongaciones del nervio óptico) que terminan en unas pequeñas estructuras denominadas conos y bastones muy sensibles a la luz. Existe un punto en la retina, llamado fóvea, alrededor del cual hay una zona que sólo tiene conos (para ver el color). Durante el día la fóvea es la parte más sensible de la retina y sobre ella se forma la imagen del objeto que se observa. Los millones de nervios que van al cerebro se combinan para formar un nervio óptico que sale de la retina por un punto que no contiene células receptoras, esto es el llamado punto ciego. La córnea refracta los rayos luminosos y el cristalino actúa como ajuste para enfocar objetos situados a diferentes distancias. De esto se encargan los músculos ciliares que modifican la curvatura de la lente y cambian su potencia. Para enfocar un objeto que está próximo, es decir, para que la

imagen se forme en la retina, los músculos ciliares se contraen, y el grosor del cristalino aumenta, acortando la distancia focal imagen. Por el contrario si el objeto está distante los músculos ciliares se relajan y la lente adelgaza. Este ajuste se denomina acomodación o adaptación.

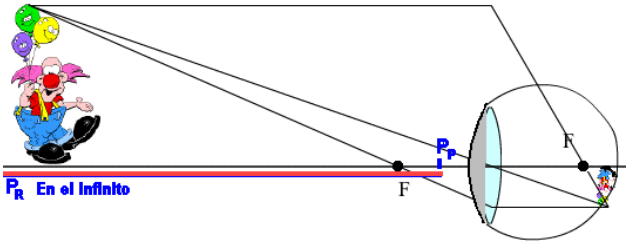


Figura A.2.- Punto Remoto.

El ojo sano y normal ve los objetos situados en el infinito sin acomodación enfocados en la retina. Esto quiere decir que el foco está en la retina y el llamado punto remoto (Pr) está en el infinito (figura A.2). Se llama punto remoto la distancia máxima a la que puede estar situado un objeto para que una persona lo distinga claramente y punto próximo a la distancia mínima. Un ojo normal será el que tiene un punto próximo a una distancia "d" de 25 cm, (para un niño puede ser de 10 cm) y un punto remoto situado en el infinito. Si no cumple estos requisitos el ojo tiene algún defecto [LW97].

El ojo es un sistema óptico que concentra y logra enfocar a la retina los rayos que salen divergentes de un objeto (de otro modo los rayos salientes de un punto no podrían recogerse sobre una pantalla para dar su imagen). En ella se puede ver que cuando el objeto se sitúa en cualquier espacio entre el punto remoto y el punto próximo la imagen se forma en la retina del ojo normal.

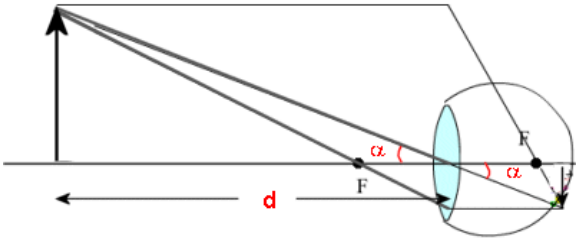


Figura A.3.- Objetos situados cercanamente.

Si un objeto está situado en el punto próximo del ojo, se ve del mayor tamaño y bajo el mayor ángulo que es posible verlo a simple vista.

A.2 La cámara fotográfica

A.2.1 Historia

Este instrumento fue descubierto por Leonardo da Vinci (1452 - 1519), quien realizó este descubrimiento cuando se encontraba en una habitación oscura protegiéndose del intenso sol de verano cuando en la pared se observaba un paisaje idéntico al exterior pero invertido. Éste fue el nacimiento de la primera idea de la cámara oscura que más tarde se transformaría en la cámara corriente fotográfica, figura A.4. A inicios del siglo XVI el árabe Ibnol Haitham estudió los eclipses solares y los de la luna. Consiguió pasar por un agujero pequeño los rayos luminosos emitidos por el sol y reflejados por la luna. Estos fueron proyectados en la pared de la habitación oscura. Este principio fue utilizado en los siglos XVII y XVIII para dibujar edificaciones y paisajes, su reproducción se lo realizaba en la parte interior de una tienda de campaña como cámara oscura. Después en el año de 1893 el Francés Daguerre empleó placas de cobre recubiertas de yoduro de plata, material sensible a la luz, que dejaba impreso el objeto observado en las placas. Sin embargo, el tipo de impresión en este material tenía un gran inconveniente que las fotografías tenían de ser preparadas con anterioridad y reveladas inmediatamente después de la exposición.

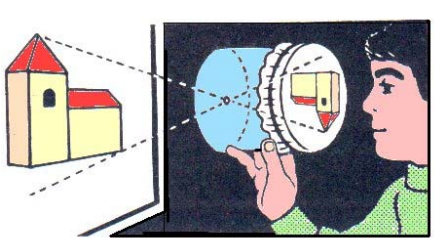


Figura A.4.- Camara obscura [WEB02].

Después de poco tiempo aparece un nuevo método descubierto por George Eastman que consistió en aplicar una placa sensible sobre una cinta flexible de celuloide de manera que los negativos obtenidos podían ser almacenados en rollos sin que estos pudieran dañarse. En el año de 1907 el científico Lumiere introdujo una nueva técnica en el comercio las primeras cámaras fotográficas para obtener fotos en colores, pero la verdadera fotografía a color apareció en 1935 cuando la compañía Kodak y Agfa produjeron fotografías con emulsión en tres capas y a todo color.

A.2.2 Elementos de la cámara fotográfica

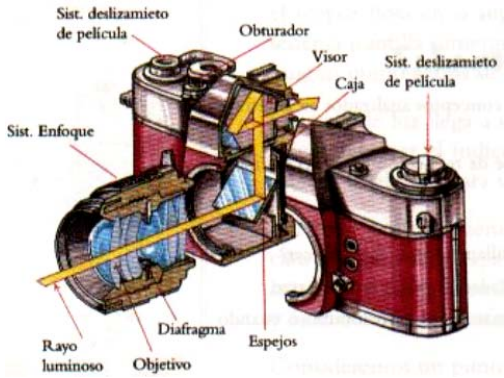


Figura A.5.- Partes de la cámara [WEB03].

A continuación se describen las partes principales de la cámara, basándose en la figura A.5:

- *Objetivo:* sistema óptico compuesto por varias lentes, que canaliza la luz que reflejan los objetos situados ante él.
- *Obturador:* sistema mecánico o electrónico que permite el paso de la luz a través del sistema óptico durante un tiempo determinado.
- *Diafragma:* sistema mecánico o electrónico que gradúa la mayor o menor intensidad de luz que debe pasar durante el tiempo que está abierto el obturador.
- *Sistema de enfoque:* gradúa la posición del objetivo, para que la imagen se forme totalmente donde está la placa sensible.
- *Sistema de deslizamiento de la película:* sistema que permite desplazar una nueva película antes de cada toma
- *Visor:* sistema óptico que permite encuadrar el campo visual que ha de ser fotografiado.
- *Caja:* estuche hermético a la luz y de color contiene todos los elementos anteriores y constituye el cuerpo de la cámara.

A.3 El ojo y la cámara fotográfica

A.3.1 Comparación entre el ojo y la cámara fotográfica

Se puede comparar el ojo con una cámara fotográfica ya que ambas estructuras tienen amplias semejanzas (figura A.6). La lente de la cámara y la córnea del ojo cumplen objetivos semejantes. Ambas son lentes positivas cuya función es la de hacer que los rayos de luz que inciden en ellas enfoquen en un solo

punto, película fotográfica o retina respectivamente. Para que córnea y lente trabajen en forma óptima deben ser perfectamente transparentes y tener las curvaturas adecuadas.

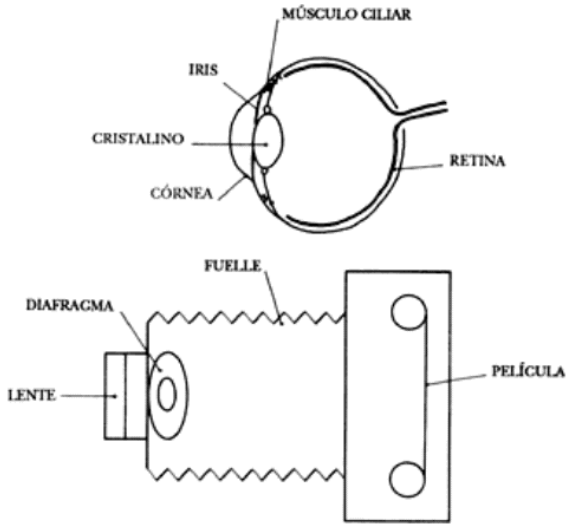


Figura A.6.- Ojo vs cámara fotográfica [WEB04].

De no ser así, la imagen proporcionada será defectuosa o no enfocará en el sitio debido. Detrás de la lente fotográfica se halla el diafragma, que es un dispositivo que regula la cantidad de luz que debe llegar a la película. A diferencia de la película fotográfica, la retina cuenta con una sensibilidad luminosa muy reducida (limitada sólo al espectro visible). En el ojo, el diafragma corresponde al iris, que es una estructura muscular perforada en su centro (pupila), y es el responsable del control de la luz que incide en la retina. Así, cuando existe poca luz ambiente, el iris se dilata creando una pupila muy grande, mientras que si la luz es intensa el iris se contrae cerrando al máximo la pupila.

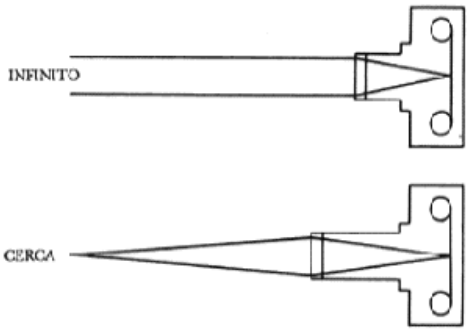


Figura A.7.- La lente Objetivo [WEB04].

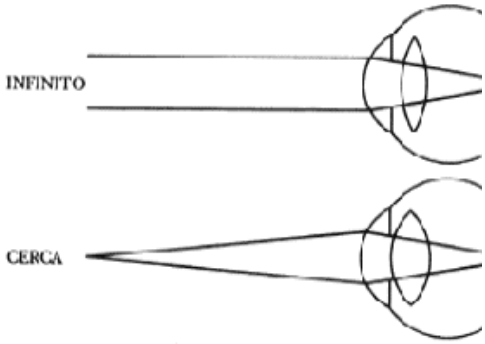


Figura A.8.- La lente cristalino[WEB04].

Al diseñar una cámara fotográfica, el poder y la posición de la lente deben calcularse de forma que los rayos paralelos de luz que incidan sobre ella enfoquen exactamente sobre la película fotográfica. Sin embargo, si el objeto se acerca a la cámara, los rayos de luz que salen de este ya no son paralelos sino divergentes, por lo que la lente objetivo, cuyo poder de refracción es fijo, ya no puede enfocarlos a la misma distancia sino detrás de la película fotográfica, tanto más lejos de ella cuanto más cerca esté el objeto por fotografiar. El sistema está entonces desenfocado. En este caso, basta con alejar la lente de la película fotográfica la distancia necesaria para que el foco caiga nuevamente sobre la película. El sistema está nuevamente enfocado, figura A.7. En las cámaras fotográficas esto se logra mediante un sistema de enfoque que permite alejar la lente de la película.

En el ojo, el proceso de enfoque existe aunque el mecanismo es distinto, figura A.8. Detrás del iris se encuentra una estructura en forma de lente biconvexa, como una lupa, llamada cristalino. Este cristalino también es transparente pero, a diferencia de la córnea, es sumamente elástico de forma que su poder refractivo es variable. En toda su periferia el cristalino está sujeto al ojo por unas fibrillas conectadas a un músculo circular. Cuando el cristalino está en reposo el sistema óptico del ojo que corresponde a la suma

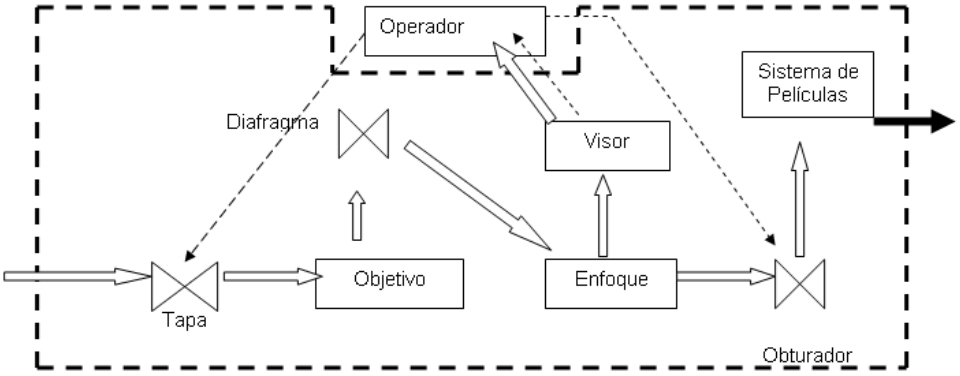
óptica de los poderes de la córnea y del cristalino hace que el ojo esté enfocado al infinito, es decir, a la visión lejana. Cuando el objeto se acerca, los rayos luminosos que llegan al ojo ya no son paralelos sino que paulatinamente se hacen cada vez más divergentes, por lo que el ojo tiene que modificar su fuerza en el músculo ciliar para poder enfocarlos en la retina.

Como ya se mencionó, en la cámara esto se obtiene alejando la lente de la película fotográfica. En el ojo, el mismo resultado se obtiene modificando las curvaturas del cristalino, es decir, haciéndolo más y más convexo conforme el objeto observado se acerca. Para ello el músculo ciliar se contrae relajando la tensión a la que está sometido el cristalino, y éste se abomba aumentando por consiguiente su poder óptico. A este fenómeno se le conoce como acomodación y es el que nos permite poder ver con nitidez los objetos cercanos.

En la cámara fotográfica la imagen del objeto llega a la película donde ocasiona cambios físicos y químicos en la emulsión, que serán tratados después en el laboratorio para fijar la imagen en el papel. En el ojo, el equivalente de la película es la retina. La retina recibe entonces la imagen en foco gracias a las propiedades ópticas de la córnea y del cristalino, con la intensidad luminosa óptima determinada por el iris. Esta imagen se fija en la retina, ocasionando cambios físicos y químicos. La gran diferencia es que esta imagen es transformada por la retina en impulsos químicos y eléctricos que viajarán posteriormente hasta los centros visuales del cerebro para hacer que la imagen sea vista por el individuo.

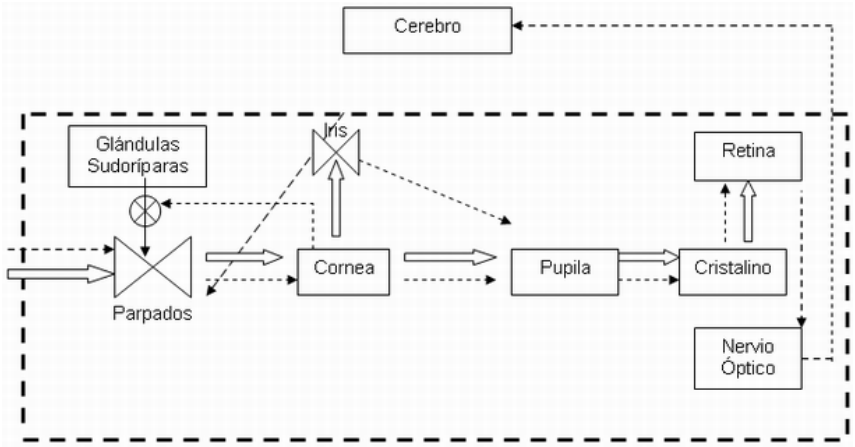
A.3.2 Diagramas de bloque correspondientes

A) La cámara fotográfica



[WEB04]

B) El Ojo



[WEB04]

A.4 Semejanzas entre el sistema visual y un sistema de vídeo

Siguiendo con las comparaciones, ahora ya no la cámara fotográfica y el ojo, sino el sistema visual completo. El hombre no ve con los ojos sino a través de los ojos. El ojo es simplemente la primera etapa de un sistema sumamente complejo. La visión es una función del sistema nervioso central, es decir es una función cerebral. Ahora en lugar de contar con una cámara fotográfica, se tiene una cámara de vídeo. El vídeo, como el cine, registra el movimiento, por lo que se parece más al ojo ya que éste además de registrar forma, tamaño y color, registra el movimiento. Con la cámara de vídeo se registra una escena familiar cualquiera, por ejemplo, la fiesta de cumpleaños. Si no se cometen errores al filmar y la cámara de vídeo funciona adecuadamente, así se tendrán registradas en la cinta las imágenes de la fiesta. Hasta aquí los hechos son semejantes a lo expuesto para la cámara fotográfica. Sin embargo, para tener acceso a la información, es decir, para ver el vídeo, se necesita de otro equipo. Analizando ahora la figura A.9. Para ver el vídeo es necesario llevar la información registrada en la cinta a una videocasetera en donde se procesa la información y se envía a un monitor (aparato de televisión) que traduce esta información en imagen. Sólo contando con el equipo completo podremos ver las imágenes de la fiesta. El sistema visual es en todo semejante al anterior.

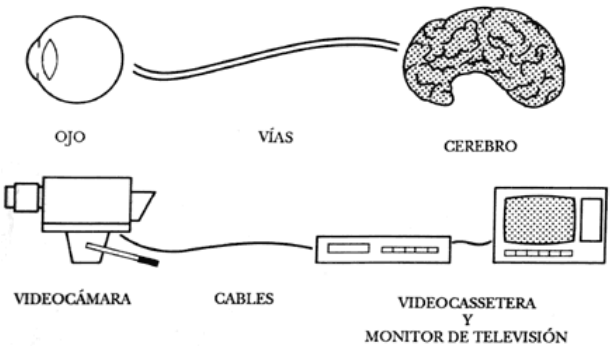


Figura A.9.- Ojo vs videocámara [WEB04].

El ojo corresponde a la cámara de vídeo. Los nervios ópticos transportan, en forma codificada, toda la información registrada en la retina a los centros analizadores del sistema nervioso en el cerebro para que el sujeto pueda ver lo que registran sus ojos. De esta forma, los centros nerviosos corresponden a la videocasetera y al monitor. El sistema visual cuenta además con otras conexiones dentro del mismo sistema nervioso que amplían enormemente sus potencialidades, permitiendo al individuo interpretar la información recibida, conectando ésta con la información de otros sistemas sensoriales, con la memoria, etcétera. Las vías visuales son entonces los nervios que parten del ojo llevando la información visual a los centros cerebrales, y los centros visuales son aquéllos localizados en la corteza occipital del cerebro y son los encargados de decodificar la información y traducirla en una percepción visual que el individuo pueda interpretar.

A.5 La cámara digital

En general, toda cámara digital tiene como objetivo capturar las imágenes, almacenarlas en la memoria interna de la cámara o en una tarjeta especial para ello y después transferirlas al ordenador; en términos técnicos, transforma los impulsos luminosos a bits, de tal manera que el ordenador al que se descarga decodifique o descifre la información [WEB05].

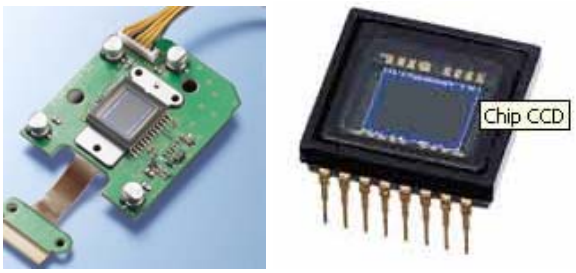


Figura A.10.- Sensor CCD [WEB06].

El elemento principal común a todas estas cámaras es un chip semiconductor sensible a la luz llamado CCD (Charge Coupled Device) o dispositivo de carga acoplada Figura A.10, creado en 1969 por Willard Boyle y George Smith de los laboratorios Bell. La película es sustituida por este dispositivo que después de filtrar los colores transforma la luz en una señal eléctrica y la almacena en la memoria de la cámara. Cuantos más valores sea capaz de recibir el CCD mejor será la calidad de las fotografías obtenidas con la cámara [WEB07].

Generalmente, el fotosensor es un CCD de tipo área (Area Array CCD), consistente en una matriz reticular de cientos de miles de células fotosensibles microscópicas (fotodiodos). A cada fotodiodo le corresponde un píxel, por lo que cuantos más fotosensores tenga el CCD, mejor será la calidad obtenida con la cámara, siendo valores habituales en las cámaras actuales 128.000 (320 x 400 píxeles de resolución) en las de gama baja, 4.200.000 (2.024 x 2.024 píxeles) en las de gama media y más de 6.000.000 en las profesionales de gama alta.

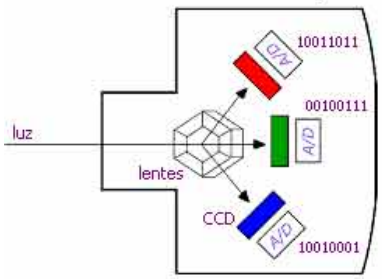


Figura A.11.- Mecanismo de una cámara digital [WEB08].

Durante el tiempo de exposición, la luz que pasa a través del juego de lentes de la cámara es dirigida a los fotosensores del CCD, figura A.11, que se encuentran cubiertos por un filtro rojo, verde o azul, encargados de dejar pasar sólo la longitud de onda correspondiente a uno de los colores básicos aditivos. Por ejemplo, el filtro rojo detiene los rayos verdes y azules, pero deja pasar el componente rojo de la luz, figura A.12.

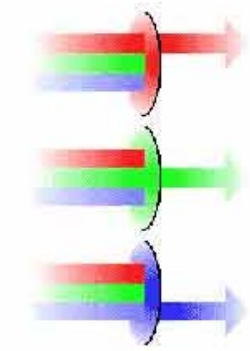


Figura A.12.- Filtros cromáticos [WEB08].

La energía luminosa filtrada es convertida entonces en cargas eléctricas, que son amplificadas y enviadas a un conversor A/D, que las transforma en información binaria de color (ceros y unos) asociada a cada uno de los píxeles de la imagen digital resultante, para pasar luego a la memoria interna de la cámara, donde es almacenada manteniendo el orden de captura, de forma que los píxeles estén dispuestos correctamente de acuerdo con el modelo fotografiado. La imagen obtenida con una cámara digital consta generalmente de millones de píxeles ordenados en líneas y columnas. Una variante mejorada del CCD es el Super CCD, desarrollado por Fujifilm, que se caracteriza por la inclusión de píxeles por interpolación para conseguir imágenes con una resolución mayor, pero incidiendo directamente en una pérdida de su calidad, figura A.13. El Super CCD dispone los píxeles octogonalmente, en forma de panel de abejas, de modo distinto al típico CCD, que lo hace rectangularmente [Fer05]. El proceso de captura puede realizarse en una o más pasadas, en cada una de las cuales se recoge información del modelo real. En caso de captura en una sola pasada, uno de cada cuatro elementos del CCD lee la información correspondiente al rojo, otro la correspondiente al verde y los dos restantes la correspondiente al azul, siendo rellenados los vacíos de información cromática que se produzcan mediante interpolación. La captura puede hacerse también en

método entrelazado, en el que el sensor de la cámara recoge información sobre la imagen procesando primero las líneas impares y luego las pares, o en el método progresivo, en el que el sensor recoge información sobre la imagen procesando las líneas de forma secuencial, una detrás de otra.

Modelo real Captura de 3600 píxeles

Figura A.13.- Píxeles capturados por una cámara digital [WEB08].

Una vez capturada la imagen, es necesario almacenarla temporalmente en la cámara hasta su descarga. Dependiendo de la marca y del modelo de la cámara se guardará la imagen digital en formatos gráficos puros, como RAW, TIF, FlashPix o Targa (TGA). Las imágenes digitales fotográficas contienen una gran cantidad de datos, por lo que generan ficheros de mucho tamaño, haciéndose necesario el uso de algún tipo de soporte que permita almacenar gran cantidad de datos en un espacio físico reducido o de mecanismos de compresión que permitan disminuir el peso del fichero gráfico (sistema habitual en las cámaras portátiles).

La compresión de la imagen es realizada por un programa específico residente en los componentes electrónicos de la cámara. En función de la calidad elegida por el usuario, esta compresión producirá un archivo de imagen JPEG de tamaño variable. Finalmente, se almacena la imagen en el soporte de almacenamiento de la cámara, normalmente tarjetas de memoria CompactFlash, figura A.14 , de las que existen los modelos CF Tipo I, de 5 mm, y CF Tipo II, de 9 mm. Las capacidades de almacenamiento más comunes varían entre 16 Mb (la más habitual en cámaras de gama baja-media), hasta 256 Mb o más.

Figura A.14.- Formatos de tarjetas [WEB09].

A.6 Cálculo de los megapixeles en el ojo humano

Se vive en un mundo audiovisual, donde hay una exagerada persecución por la calidad de la imagen, por la resolución que ofrecen diversos aparatos electrónicos, pero primero se tendría que responder qué resolución tiene el ojo, para saber si tanta resolución en dichos aparatos es suficiente. El número de receptores de nuestra retina es de un orden aproximado de 85-126 millones (80-120 millones de bastones y 5-6 millones de conos), es decir, que cada ojo tendría en torno a 100 megapixels. Pero las fibras en el nervio óptico sólo son de entre un millón hasta de millón y medio, por lo que al cerebro sólo le llega una imagen de 1 o 1.5 megapixels [CG66]. Lo anterior es una mala estimación, porque principalmente, el ojo no es una cámara. Los dos ojos no paran de parpadear y se mueven para cubrir un área mucho mayor que al campo de visión y la composición de todas esas imágenes son unidas y analizadas por el cerebro, de una forma más compleja y precisa que un simple ensamblaje de fotos. En un ambiente con luz buena, se podrían distinguir dos líneas finas si estuvieran separadas sólo por 0.6 minutos de arco (0.01 grados).Esto sí que da una equivalencia con las cámaras, que es un pixel de 0.3 minutos de arco. Si se parte de que nuestro campo visual humano es de 120 grados en horizontal y 60 grados en vertical, da un total aproximado de 576 megapixeles de datos en esa imagen que se está viendo.

Anexo B

Visión Estereoscópica

B.1 Los primeros pasos de la visión estereoscópica

Una de las primeras investigaciones sobre la visión binocular fue introducida por Charles Wheatstone [Whe38], donde se describe que cuando un objeto es visto a una gran distancia, los ejes ópticos de ambos ojos son ligeramente paralelos y cuando se dirige la mirada a dicho objeto, la perspectiva de él es vista por cada ojo de manera separada, ambas imágenes son muy parecidas y la apreciación es la misma que cuando el objeto es visto sólo por un ojo. En tal caso, no hay diferencia entre la apreciación de un objeto en relieve y su perspectiva en una superficie plana; y por tanto las representaciones gráficas de objetos distantes, son excluidas cuando estas perturben la ilusión, pero también se descartan similitudes de los objetos que se intentan representar. Pero estas similitudes no permanecen cuando el objeto es localizado, tan cerca de los ojos, que los ejes ópticos puedan converger, bajo estas circunstancias es vista por cada ojo una diferente perspectiva y estas apreciaciones son diferentes a medida que la convergencia de los ejes ópticos sea mayor.

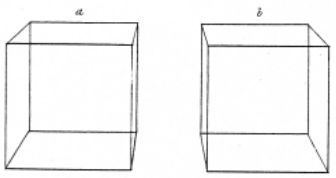


Figura B.1.- Las dos perspectivas de un cubo.

La figura B.1 representa las dos apreciaciones de un cubo; el cubo *b* es el que se ve con el ojo derecho y el cubo *a* es lo que ve el ojo izquierdo, dicha figura supone que se coloque a alrededor de siete pulgadas inmediatamente delante del espectador. Las apariencias, las cuales por este simple experimento resultan ser tan obvias, se podrían deducir fácilmente de las leyes dadas de la perspectiva; el mismo objeto en relieve, cuando es apreciado por un diferente ojo, es visto de dos puntos de vista diferentes a una distancia mutua igual a la línea de acoplamiento de los dos ojos.

B.2 La geometría del sistema visual humano

La percepción de la profundidad es una de las más importantes características de cualquier muestra de entornos que son tridimensionales. El sistema visual humano tiene una configuración que soporta dos imágenes separadas recogidas por cada ojo. Entonces el cerebro las combina en una sola, una pequeña, pero importante diferencia matemática existe entre dichas imágenes. Esta minúscula diferencia es representada por la figura B.2 [BT04]. En la figura B.3, A es el ángulo de los ojos para observar imágenes y movimiento (el movimiento solamente de ángulos agudos y obtusos, es decir de 0 a 180°), B equivale al punto en el cual la nariz interfiere con la habilidad de los ojos en ver imágenes y movimiento, mientras que C es igual a la distancia entre la línea central de las pupilas.

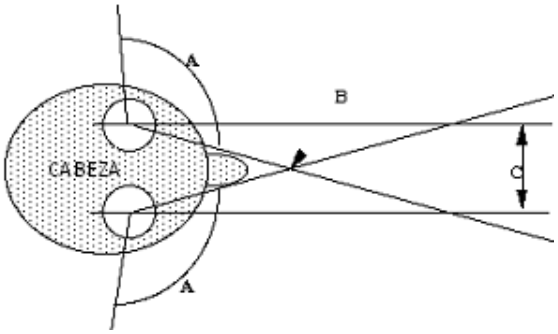


Figura B.2.- El sistema visual humano.

Esta pequeña compensación C, causa la localización de la imagen en el campo visual, por ejemplo el ángulo A del ojo izquierdo es diferente que en el ojo derecho a pesar de ser la misma imagen. Observando la Figura B.3 se aprecia como cada ojo puede ver la misma imagen u objeto de diferente forma.

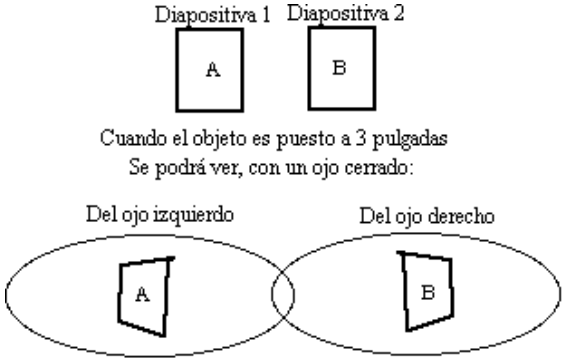


Figura B.3.- Diferencia ente el ojo derecho e izquierdo.

Este efecto agrega información de percepción de profundidad entonces, el cerebro deducirá una imagen con tres dimensiones, combinando la información obtenida por cada ojo y así formar una sola de forma tridimensional. El sistema visual humano solamente soporta esta percepción de profundidad estereoscópica en objetos que están entre 20 centímetros y 5.5 metros, después de eso el sistema visual ocupa otras señales para detectar la profundidad y la distancia. A grandes distancias, los ojos pueden detectar movimientos muy sutiles y cambios en las relaciones entre los objetos. Estos cambios son la fuente primaria de información en las señales de percepción de la distancia y profundidad para ver a distancias mayores de 5.5 metros.

Para encontrar la correspondencia, algunas restricciones en las imágenes estéreo deben ser asumidas antes, como se detalla a continuación:

- *Unicidad:* cada punto tiene a lo más uno igual en la otra imagen.
- *Similitud:* cada área con una intensidad de color definida iguala a la otra área en la otra imagen.
- *Ordenamiento:* el orden de los puntos en las dos imágenes normalmente es el mismo.
- *Continuidad:* los cambios de la disparidad varían ligeramente a lo largo de la superficie, excepto en los bordes de profundidad.
- *Restricción epipolar:* dado un punto en la imagen obtenida por un ojo, el punto similar en la imagen para el otro ojo debe trazarse en únicamente una línea recta.

B.3 Geometría Epipolar

B.3.1 Bases

La geometría epipolar es determinada por las correspondencias en los puntos. La selección y correspondencia de los puntos característicos en las dos vistas son el procedimiento estándar para recuperar la profundidad. Dicha información de profundidad puede ser evaluada por el uso de algoritmos parecidos a

triángulos. La geometría epipolar puede representarse, como muestra la figura B.4, en la parte superior de la vista. En dicha ilustración, tanto la distancia base (T) como la longitud del enfoque (f) de ambas cámaras son conocidas. La proyección de perspectiva Pl y Pr de la línea epipolar es cambiada desde sus centros por las distancias xl y xr respectivamente. La distancia desde el objetivo a la línea base (Z) puede ser determinada por comparación de la similitud de los triángulos P-Fl-Fr y P-Pl-Pr. Las ecuaciones B.1, B.2 y B.3 presentan la evaluación de distancia del objeto.

$$\frac{T + x_l - x_r}{Z - f} = \frac{T}{Z} \qquad \text{……..(B.1)}$$

Entonces:

$$Z = f\frac{T}{x_r - x_l} = f\frac{T}{d} \qquad \text{……..(B.2)}$$

Donde: *d* es la disparidad.

$$d = x_r - x_l \qquad \text{……..(B.3)}$$

El punto de proyección puede ser computado por cada vista y cada punto de la imagen que sean iguales, donde puedan ser proyectados de forma inversa para dar una estructura tridimensional.

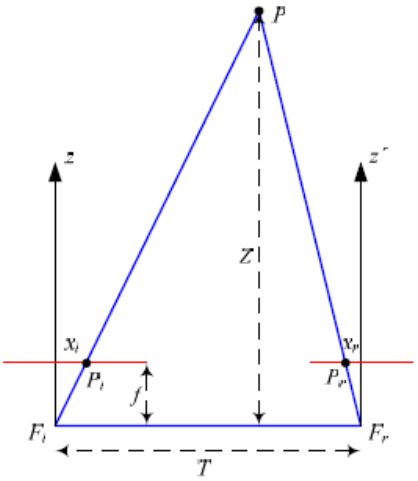


Figura B.4.- La disparidad es el desplazamiento entre las posiciones de dos puntos proyectados en el plano de la imagen.

Los vectores $P_l=[X_l,Y_l,Z_l]^T$ y $P_r=[X_r,Y_r,Z_r]^T$ se refieren al mismo punto 3D, P, como vectores en los sistemas de referencia de la cámara izquierda y derecha respectivamente. Los vectores $x_l=[x_l,y_l,z_l]^T$ y $x_r=[x_r,y_r,z_r]^T$ definen las proyecciones del punto P en la imagen izquierda y derecha respectivamente y están expresados en su correspondiente sistema de referencia. Evidentemente, para todos los puntos de las imágenes se tiene $z_l = x_l$ ó $z_r = x_r$ de acuerdo a la imagen que sea.

Los sistemas de referencia de las cámaras izquierda y derecha están relacionados a través de los parámetros extrínsecos. Estos definen una transformación rígida en el espacio 3D definida por un vector de traslación de la ecuación B.4.

$$T = F_r - F_l \qquad \text{……..(B.4)}$$

y una matriz de rotación R. dado un punto P en el espacio la relación entre P_l y P_r expresado en la ecuación B.5.

$$P_r = R(P_l - T) \qquad \text{……..…(B.5)}$$

El nombre de geometría epipolar es debido a que los puntos en los cuales la recta que une los centros de proyección de las cámaras corta a los planos de proyección se llaman epipolos. Se denotará por e_l y e_r el epipolo izquierdo y derecho respectivamente. Por construcción ambos epipolos representan la proyección en su correspondiente plano la imagen del centro de proyección de la otra cámara. En el caso de que uno de los planos imagen sea paralelo a la recta que une los centros de proyección, el epipolo de ese plano estará situado en el infinito. La relación entre un punto del espacio 3D y su proyección se describe por las ecuaciones usuales de proyección de perspectiva, que en forma vectorial se escriben como en las ecuación B.6 y B.7.

$$x_l = \frac{f}{Z_l} P_l \qquad \text{……..……..……..(B.6)}$$

$$x_r = \frac{f}{Z_r} P_r \qquad \text{……..……..……..(B.7)}$$

La importancia práctica de la geometría epipolar arranca del hecho que el plano identificado por P, x_l, x_r llamado plano epipolar, intercepta cada imagen en una línea llamada línea epipolar. Considerando la tercia P, x_l y x_r. Dado x_l P puede caer en cualquier punto del rayo definido por F_l y x_l. Pero dado que la imagen de este rayo en la imagen derecha es la línea epipolar a través del punto correspondiente x_r dicho punto debe estar sobre la línea epipolar. Esta correspondencia establece una aplicación entre puntos de la imagen izquierda y rectas de la imagen derecha y viceversa. Una consecuencia de esta correspondencia es, que dado que todos los rayos pasan por construcción por el centro de proyección, todas las rectas epipolares deben pasar por el epipolo. Por tanto si se determina la aplicación entre puntos de la imagen izquierda (derecha) y las rectas epipolares de la imagen derecha (izquierda), se restringe la búsqueda para el emparejamiento de x_l a lo largo de la línea epipolar correspondiente. Así pues la búsqueda de las correspondencias se reduce a un problema 1D. Alternativamente este conocimiento también se usa para verificar sí una potencial pareja de puntos correspondientes, lo son de verdad o no. Esta técnica es normalmente una de las más efectivas para detectar las posibles falsas correspondencias debidas a oclusión.

B.3.2 Matrices

MATRIZ ESENCIAL

La ecuación del plano epipolar a través de P puede escribirse como la condición co-planar de los vectores P_l,T y P_l - T, o $(P_l - T)^T$ (T x P_l) = 0. Usando la relación que liga a los vectores P_l y P_r se obtiene $(R^T P_r)^T$ (T_x P_l) = 0. Teniendo en cuenta que el producto vectorial de dos vectores se puede escribir como la multiplicación de una matriz antisimétrica por un vector, se tiene $T_x P_i = [T]_x P_i = SP_i$ donde:

$$[T]_X = S = \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix}$$

Y con ello se obtiene que $P_r^T E P_l = 0$ con E=RS=0.

Se observa por construcción E siempre tiene rango igual a 2 y se puede probar que su dos autovalores distintos de cero son iguales. La matriz E se denomina la matriz esencial y establece una unión natural entre la restricción epipolar y los parámetros extrínsecos del sistema estéreo. Ahora sí, se consideran las ecuaciones vectoriales de la perspectiva y se sustituye en la ecuación anterior se obtiene la ecuación que liga las proyecciones del punto P en ambas planos imagen y se divide por $Z_l Z_r$ se obtiene la ecuación B.8.

$$x_r^T E x_l = 0 \qquad \text{……..(B.8)}$$

El vector que representa $u_r=Ex_l$ puede ser interpretado como el vector director de la recta en que se proyecta, sobre la imagen derecha, el rayo definido por F_lP. Así pues, y a través de la matriz esencial se establece una aplicación entre los puntos de una imagen y las rectas epipolares de la otra (recordar cómo se han definido las rectas epipolares asociada a un punto P). Usando la notación vectorial introducida, la ecuación anterior se escribe como $x_r^T u_r = 0$, que establece la condición de incidencia de que uno de los puntos siempre se encuentra sobre la línea epipolar definida por el otro.

Hasta ahora se ha trabajado en coordenadas de los sistemas de referencia asociados a las cámaras, sin embargo, cuando se calculan los puntos proyección estos están medidos en términos de píxeles, por tanto para aprovechar la matriz esencial será necesario conocer la transformación desde coordenadas de la cámara a coordenadas de píxeles, es decir, conocer los parámetros intrínsecos. Esta restricción puede ser eliminada pero a costa de pagar un precio sobre la información recuperada.

MATRIZ FUNDAMENTAL

Ahora se expone que la aplicación entre puntos y líneas, lo cual se logra a partir de puntos correspondientes solamente, sin necesidad de información a priori sobre el sistema estéreo.

Sean K_l y K_r las matrices de los parámetros intrínsecos de las cámaras izquierda y derecha respectivamente. Las ecuaciones B.9 y B.10 denotan a $\overline{x_l}$ y $\overline{x_r}$ como los puntos en coordenadas píxel correspondientes a p_l y p_r respectivamente.

$$x_l = K_l^{-1}\overline{x_l} \qquad \text{……..(B.9)}$$

$$x_r = K_r^{-1}\overline{x_r} \qquad \text{……..(B.10)}$$

y sustituyendo en la ecuación de la matriz esencial da como resultado que $\overline{x_l}^T F \overline{x_r} = 0$ donde F se denomina la matriz fundamental y $F = K_r^{-1} E K_l^{-1} = K_r^{-T} R S K_l^{-1}$. Al igual que con $Ex_i, F\overline{x_i}$ puede ser interpretado como el vector de la recta epipolar proyectiva correspondiente al punto $\overline{x_i}, u_r = F\overline{x_i}$. La mayor diferencia entre las ecuaciones en términos de la matriz esencial y la matriz fundamental es que la matriz esencial está establecida en términos de vectores definidos en los sistemas de referencia de las cámaras, mientras que la matriz fundamental está definida en términos de vectores definidos en términos de coordenadas píxeles de los planos de proyección. Consecuentemente, si se estima la matriz fundamental a partir de puntos en correspondencia en coordenadas píxel se puede reconstruir la geometría epipolar sin absolutamente ninguna información sobre los parámetros intrínsecos o extrínsecos.

Todo lo anterior indica que es posible establecer la correspondencia entre los puntos de una imagen y sus correspondientes líneas epipolares sin ningún conocimiento a priori de los parámetros del sistema estéreo. Tanto la matriz fundamental como la matriz esencial no son matrices de rango completo, siendo de rango igual a 2 ya que las matrices de los parámetros intrínsecos son de rango completo. La matriz fundamental codifica información sobre los parámetros tanto intrínsecos como extrínsecos.

Una consecuencia muy importante de las anteriores propiedades es que si se conoce la calibración de un sistema estéreo o de una cámara, (parámetros intrínsecos y parámetros extrínsecos) se sabe toda la información necesaria para obtener los cálculos de la geometría epipolar y la aplicación que liga puntos de una imagen con sus correspondientes rectas epipolares.

B.3.3 Localización de los epipolos a partir de la matriz F

Estableciendo la relación que liga a la matriz F con los epipolos de los planos retinales. Sea $\overline{e_l}$, el epipólo del plano izquierdo. Dado que es un punto por el que pasan todas las rectas epipolares asociadas a los puntos de la imagen derecha, se verifica que $\overline{x_r}^{-T} F \overline{e_l} = 0$ para todo $\overline{x_r}$. Pero dado que F no es idénticamente nula, esto tan solo será posible si se verifica que $F\overline{e_l} = 0$. Dado que F es de rango igual a 2, el epipolo $\overline{e_l}$ se puede por tanto calcular como el autovector asociado al autovalor nulo de la matriz F. Haciendo un razonamiento similar se demuestra que el epipolo de la imagen derecha es el autovector asociado al autovalor nulo de la matriz F^{T}, $F^T \overline{e_r} = 0$ [HZ04].

Las consideraciones para la matriz E son similares a las realizadas para la matriz F. La figura B.5 muestra las líneas epipolares y el epipolo de las imágenes en distintas posiciones de la cámara respecto del objeto:

a) la parte superior de la figura (a) muestra la situación relativa de las cámaras. La parte inferior muestra las imágenes de cada cámara y líneas epipolares conjugada (b) y (c).

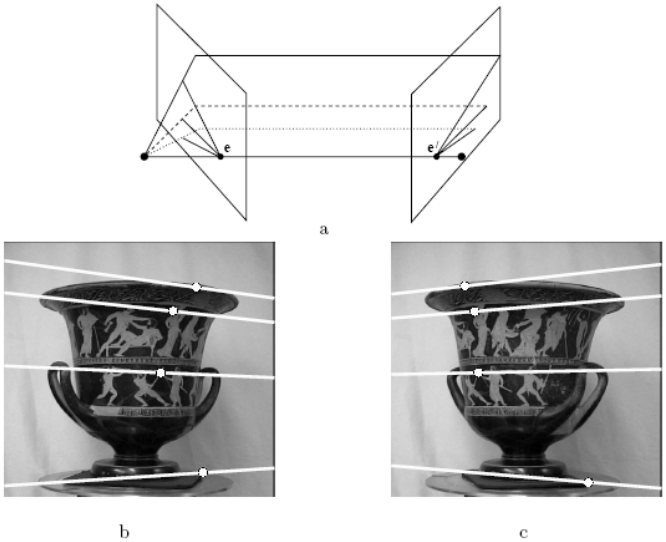


Figura B.5.- Dos cámaras que convergen [WEB10].

La imagen de la figura B.6 muestra dos planos prácticamente paralelos de la misma escena. Obsérvese que las líneas epipolares son paralelas, lo que indica que los epipolos están en el infinito.

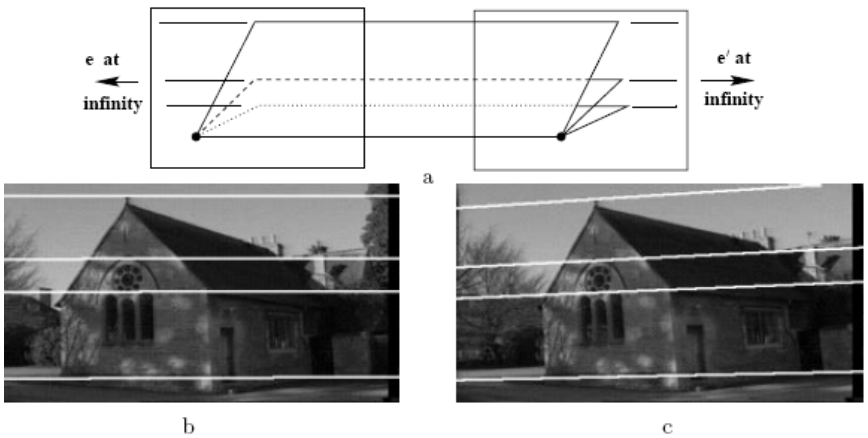


Figura B.6.- Paralelaje de la misma escena [WEB10].

B.3.4 Matrices fundamentales asociadas

A LA TRASLACIÓN PURA

Al considerar este caso se supone que la cámara está quieta y el mundo moviéndose en una traslación –t. En esta situación, los puntos del espacio se mueven siguiendo líneas rectas paralelas a t cuyo punto de intersección será el punto de anulación v en la dirección t.

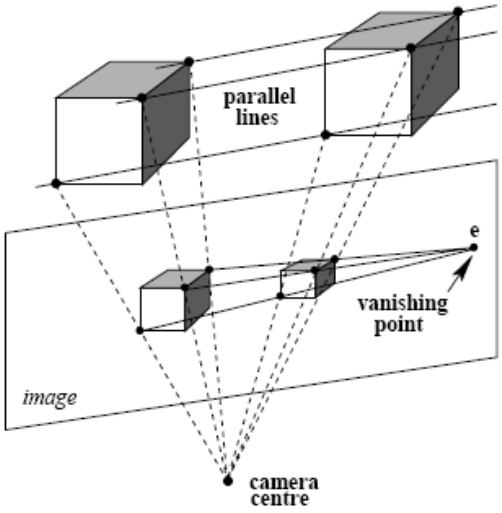


Figura B.7.- Translación pura [WEB10].

Suponiendo que el movimiento de la cámara es de pura traslación sin rotación y sin cambio de parámetros internos. Las cámaras son P=K[I |0] y P'=K[I | t]. Entonces la matriz fundamental será:

$$F = [e']_x KK^{-1} = [e']_x$$

Si la cámara se traslada paralela el eje x, entonces $e'=(1,0,0)^T$, por tanto:

$$F = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

y la relación $x'^T Fx=0$ se reduce a y=y', es decir las líneas epipolares se corresponden con las filas de la imagen.

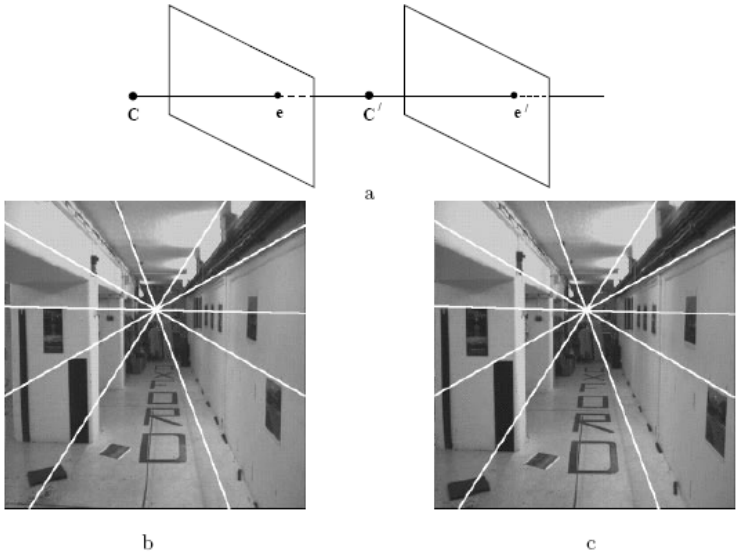


Figura B.8.- Ejemplo de translación pura [WEB10].

Si el punto imagen x se normaliza a $\tilde{x} = (x, y, 1)^T$ entonces de la ecuación de la proyección x=PX=K[I | 0]X calcula las coordenadas no-homogéneas (X,Y,Z) de los puntos del espacio como $(X, Y, Z)^T = K^{-1}\tilde{x} = \frac{K^{-1}x}{Z}$ siendo Z la profundidad del punto X (distancia medida desde el punto al centro de la cámara a lo largo del eje principal). Se sigue entonces de la proyección del mismo punto X en la otra cámara, x'=P'X=K[I | t] X que la aplicación de un punto x de una imagen a una punto x' lo que da como resultado la ecuación B.11.

$$x' = x + \frac{Kt}{Z} \qquad \text{.....(B.11)}$$

La ecuación B.11 muestra que el punto comienza en x y se mueve a lo largo de la línea definida por x y el epipolo e=e'=v . La extensión del movimiento depende del vector de traslación t y de la profundidad del punto Z, por tanto puntos más cercanos a la cámara aparecen moverse más rápidos que puntos más alejados. En este caso de traslación pura F = $[e']_x$ es antisimétrica y tiene por tanto solamente dos grados de libertad que corresponden a la posición del epipolo. La línea epipolar de x es Fx=$[e]_x$x= e x x, es decir, x pertenece a su línea epipolar, luego se da un caso especial de auto-epipolaridad que no se verifica en un movimiento general.

AL MOVIMIENTO GENERAL

El caso anterior de traslación pura una idea adicional sobre el movimiento general. Ya que dadas dos cámaras arbitrarias es posible mostrar que es posible definir una transformación proyectiva H tal que aplicada a la primera deje ambas cámaras alineadas (rotación + ajuste de las calibraciones de ambas cámaras). En ese caso la matriz fundamental de las cámaras alineadas sería F = $[e']_x$. Consecuentemente y deshaciendo la transformación H la matriz fundamental de las cámaras iníciales es F = $[e']_x$ H. Si las cámaras tienen matrices P=K[I | 0] y P'=K'[R |t] entonces la transformación sería H= K'R K^{-1} =H_∞ y F = $[e']_x$ H_∞.

Al igual que antes se calculan las coordenadas 3D de los puntos $(X, Y, Z)^T = K^{-1}\tilde{x} = \frac{K^{-1}x}{Z}$. Y usando la ecuación de proyección en la otra cámara se obtiene la ecuación B.12.

$$x' = K'RK^{-1}x + \frac{K't}{Z} = H_\infty x + \frac{K't}{Z} = H_\infty x + \frac{e'}{Z} \qquad \text{.....(B.12)}$$

La expresión de la ecuación B.12 muestra que el movimiento de un punto en una imagen por un movimiento general de una cámara se compone de dos partes: una que sólo depende del giro y de los parámetros de calibración y otro que depende de la traslación efectuada y de la profundidad del punto en la escena.

AL MOVIMIENTO PLANO PURO

En este caso es de destacar que la dirección del movimiento es siempre perpendicular a la normal al plano, lo cual induce una restricción sobre la matriz fundamental ya que es posible demostrar que la parte simétrica $F_s = \frac{\left(F + F^T\right)}{2}$ de F tiene rango 2 lo cual le resta un grado de libertad dejándola sólo con 6 grados de libertad.