Método para detección y seguimiento de objetos con aplicaciones en Realidad Aumentada

Christian Nicolás Pfarher

Punto de control N° 4:

Pre-procesamiento para lograr posibles mejoras de velocidad, en tiempo de ejecución

Análisis homográfico, proyecciones y transformaciones

Superposición de contenido sobre el objeto detectado

Director
Dr. Enrique Marcelo Albornoz

Codirector
Dr. César Martínez

29 de junio de 2012

Ingeniería Informática
Facultad de Ingeniería y Ciencias Hídricas
UNIVERSIDAD NACIONAL DEL LITORAL

Índice

1. Pre-procesamiento para lograr posibles mejoras de velocidad, en tiempo de ejecución

En esta sección, se plantearán algunas herramientas como posibles alternativas a aplicar con el objetivo de lograr mejoras de velocidad en tiempo de ejecución.

Como es sabido, tanto el uso del método SURF para la extracción de características como la búsqueda de coincidencias entre imágenes, conllevan un gran tiempo de procesamiento, el cual se convierte en un factor clave a la hora de lograr mayor fluidez en la reproducción del flujo de video [Lag11, BL97, FBF77, LMGY04]. Para reducir este problema, se recurre a técnicas para detectar la parte cambiante de la imagen, de tal forma de aplicar el procesamiento solo en una región de la imagen capturada en vez de en su totalidad de forma de reducir los datos a procesar. Para ello, se utilizan diferentes herramientas de procesamiento de imágenes en el siguiente orden: conversión a escala de grises, diferencia de imágenes, umbral, erosión, dilatación y bounding box. Estas herramientas, fueron elegidas debido a su grado de simpleza y bajo tiempo de procesamiento que requieren para llevarse a cabo, de forma tal de influenciar en la menor medida posible el tiempo de proceso total del algoritmo.

1.1. Conversión a escala de grises

La conversión a escala de grises, consiste en convertir la imagen de un espacio de color a otro, en este caso se convierte del espacio de colores RGB o BGR a escala de grises. Para el cálculo del valor de gris se usa la fórmula perceptualmente ponderada $Y = (0{,}299)R + (0{,}587)G + 0{,}114B$ donde R, G y B representan los canales rojo, verde y azul de la imagen respectivamente.

Este proceso es llevado a cabo mediante la función provista por la librería OpenCV $cvCvtColor(constCvArr * src, CvArr * dst, intcode)$ con el parámetro $code = CV\_RGB2GRAY$.

1.2. Diferencia de imágenes

La diferencia de imágenes se aplica con el objetivo de lograr distinguir el cambio en la imagen cuadro a cuadro. Consiste en calcular el valor absoluto de la diferencia entre el cuadro F del flujo de video capturado en el tiempo t (actual) y el capturado en $t-1$ (anterior), es decir: $D = |F_t - F_{t-1}|$ (pixel a pixel). El resultado de la operación, es una imagen D en la que se observa la parte que ha cambiado de una imagen a otra.

1.3. Threshold

A partir de la imagen diferencia D obtenida anteriormente, se procede con la aplicación de un umbral binario de la forma (1) para remover ruido.

Se asigna el valor $valorMaximo = 255$ a los píxeles que superen el umbral con un valor definido empíricamente ($u = 50$) y 0 a los demás píxeles, obteniéndose como resultado una nueva imagen I.

$$I = \begin{cases} x & \text{si } valorMaximo > u, \\ 0 & \text{en otro caso} \end{cases} \tag{1}$$

1.4. Erosión y dilatación

Las erosión y dilatación son dos de las principales operaciones morfológicas usadas en variedad de contextos como remoción de ruido, aislación/unión de elementos, etc.

En este contexto, estas operaciones serán aplicadas para eliminar puntos aislados o ruido que no haya sido removido mediante el umbral descripto anteriormente en 1.3, de forma que el Bounding Box sea detectado correctamente en una próxima etapa.

Tanto la **Dilatación** como la **Erosión** consisten en la convolución de una imagen (o una parte de ella) con un kernel que puede ser de diferentes formas o tamaños y que posee un "punto de anclaje" también llamado punto central del kernel.

La *erosión*, reemplaza al píxel actual con el valor mínimo de la vecindad definida, mientras que la *dilatación* es la operación complementaria a la anterior, es decir, que reemplaza el píxel actual con el valor máximo del conjunto de píxeles vecinos. Así, en el caso de una imagen binaria (sólo contiene píxeles negros (0) y blancos (255)), cada píxel es reemplazado por uno negro o blanco. En una imagen erosionada, el tamaño de los objetos se ve reducido y el ruido es eliminado. En el caso de la dilatación, los objetos crecen en su tamaño y algunos de los espacios dentro de ellos son rellenados.

1.5. Bounding Box

Luego del procesamiento anterior, es el momento de detectar el rectángulo más pequeño que encierra completamente todos los puntos cuyo valor de píxel no es completamente 0, es decir, píxeles no negros. A esto se le llama Bounding Box y puede ser utilizado para descartar regiones de la imagen que no necesitamos procesar en un algoritmo. Debido que anteriormente hemos eliminado el ruido mediante las operaciones de umbral, dilatación y erosión, nos aseguramos de esta forma el poder encontrar un único Bounding Box válido. Este contedrá la parte cambiante de la imagen sobre la que detectaremos características y haremos el posterior proceso de correspondencia y detección de la ubicación del objetivo a detectar.

2. Análisis homográfico, proyecciones y transformaciones

Hasta el momento, se han mencionado diversas técnicas que se aplican a la imagen para detectar características distinguibles de las mismas (puntos claves y vectores descriptores), de forma de poder identificarla en un contexto específico. A continuación se expondrá la forma por la cual se detecta si el objeto a detectar está presente en el flujo de video capturado, además de identificar su posición respecto a la cámara. Para ello, es necesario introducir algunos conceptos básicos de formación de la imagen que se explicarán a continuación.

2.1. Introducción

El principal componente para la visión en una escena es la luz. La misma proviene como un rayo que sale de una fuente (por ejemplo: el sol o una lámpara) y viaja a través del espacio hasta intersectar un objeto. Aquí, parte de la luz es absorbida y otra reflejada (percibida como el color) la cual es captada por los ojos (o cámara) y recogida en la retina (o imagen).

Existen diversos dispositivos para obtener imágenes, uno de los más comunes, es la cámara digital. Ésta captura la escena mediante la proyección de luz en un sensor a través del lente de la cámara. El hecho de que la imagen se forma a través de la proyección de la escena 3D en un plano 2D, implica la existencia de importantes relaciones entre la escena y su imagen, y es la geometría proyectiva la herramienta usada para describir y caracterizar en términos matemáticos este proceso de formación de la imagen [BK08, Lag11].

Para entender un poco más el proceso, existe un modelo teórico útil denominado **modelo de caja oscura**, del inglés: **"pinhole camera model"** [HZ04, BK08] con el cual se puede entender la geometría básica en la proyección de rayos y que será expuesto en 2.2. Sin embargo, un modelo de este tipo no puede ser aplicado realmente para obtener imágenes en la vida real, ya que el mismo no puede recoger la suficiente cantidad de luz en una exposición rápida. Es por esto que las cámaras usan lentes (los cuales introducen distorsiones no deseables) para obtener más luz de la disponible para un punto dado. Estas desviaciones provocadas por la lente, pueden ser corregidas matemáticamente a través de la calibración de la cámara, técnica mediante la cual se puede obtener tanto el modelo geométrico de la cámara como el modelo de distorsión de la lente, los cuales pasan a formar parte de los parámetros intrínsecos.

2.2. Formación de la imagen

El proceso de formación de la imagen no ha cambiado desde los comienzos de la fotografía [Lag11]. La luz proveniente de la escena observada, es capturada por la cámara a través de la apertura frontal y los rayos de luz impactan en el plano de la imagen (o sensor de imagen) localizado por detrás de la cámara, en donde un lente es usado para concentrar los rayos provenientes de los diferentes elementos de la escena. Este proceso, se ilustra en la Fig. 1, donde *do* es la distancia de la lente al objeto observado, *di* es la distancia de la lente al plano de la imagen y f es denominada **distancia focal**. Estas variables, se encuentran relacionadas mediante la ecuación de la lente (2).

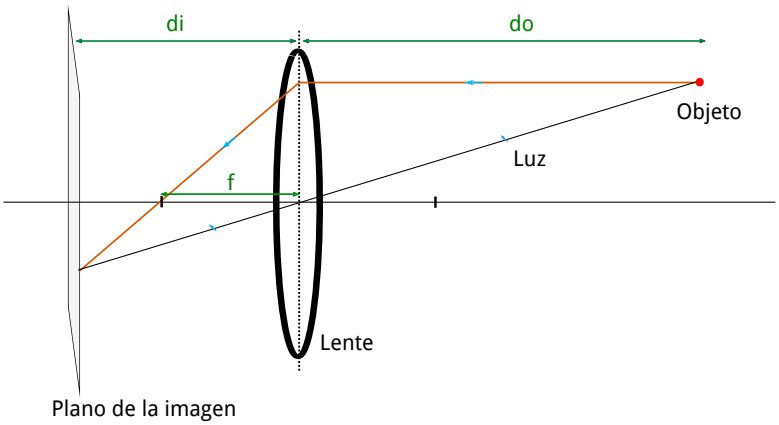

Figura 1: Formación de la imagen. Adaptada de [Lag11].

$$\frac{1}{f} = \frac{1}{do} + \frac{1}{di} \tag{2}$$

En visión computacional, este modelo de cámara puede ser simplificado mediante ciertas consideraciones:

- Se asume nulo el efecto de la lente (distorsión),
- Se considera una cámara con una apertura infinitesimal: de esta forma sólo se considera el rayo central,
- Se asume $do >> di$: esto significa que el plano de la imagen está en la posición de la distancia focal (enfocado),
- De la geometría del sistema, puede observarse que el objeto resulta invertido en el plano de la imagen: se puede obtener una imagen idéntica pero no invertida, simplemente posicionando el plano de la imagen en frente de la lente (aunque esto no es físicamente posible, es completamente equivalente desde un punto de vista matemático).

Este modelo simplificado, es conocido como "modelo de cámara oscura" y se representa en la Fig. 2

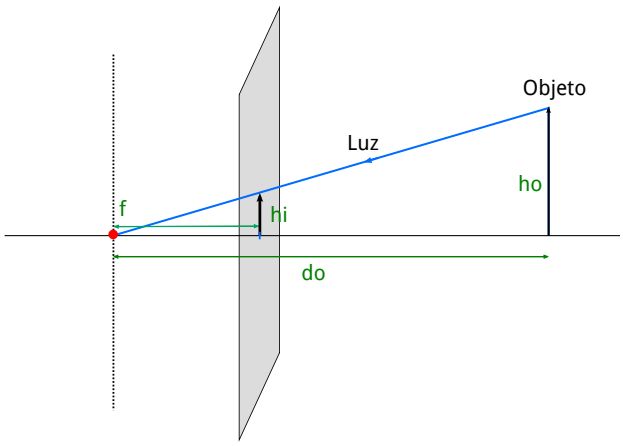

Figura 2: Modelo pinhole de la cámara. Adaptada de [Lag11].

Usando la ley de similitud de triángulos y observando el modelo mencionado, se puede derivar la ecuación de proyección (3) de la cual se desprende que: el tamaño *hi* de un objeto de altura *ho* en el plano imagen, resulta inversamente proporcional a la distancia *do* entre la cámara y el objeto. Esta relación, es fácilmente observable: cuando se fotografía un objeto cercano a la cámara, el mismo aparece de mayor tamaño en el plano de la imagen; por el contrario si el objeto es alejado de la misma, éste aparece de menores dimensiones en el plano de la imagen. De esta manera se puede obtener la posición de un punto 3D de la escena, en el plano imagen.

$$hi = f\frac{ho}{do} \tag{3}$$

2.3. Calibración de la cámara

De la sección 2.2, se vio que los parámetros esenciales de la cámara son la distancia focal y el tamaño del plano de la imagen. Cuando se trata de imágenes digitales, la cantidad de píxeles en el plano de la imagen resulta ser otro parámetro importante a determinar con el objetivo de calcular la posición de un punto de la escena en el plano de la imagen. Si consideramos la línea ortogonal al plano de la imagen proveniente del punto focal, necesitamos saber en que píxel esta línea perfora el plano de la imagen. Este punto es llamado **"punto principal"**. Si bien resulta lógico asumir que este punto está en el centro del plano de la imagen, esto sólo se puede aseverar en condiciones ideales, ya que en la práctica el mismo se encuentra desplazado dependiendo de la precisión de fabricación de la cámara. Es por ello que se necesita un proceso de calibración de la misma.

La calibración de la cámara es el proceso, por el cual son obtenidos diferentes parámetros tales como la matriz de esta y los parámetros de distorsión (ignorados en el modelo de cámara oscura mencionado anteriormente) [BK08, Lag11].

Para explicar los resultados de la calibración de la cámara, debemos referirnos a lo descripto en el modelo de cámara oscura introducido en la sección 2.2, para así entender la relación entre un punto 3D de la escena P en la posición (X, Y, Z) y su correspondiente en el plano de la imagen p en la posición (x, y), especificado en coordenadas de píxeles.

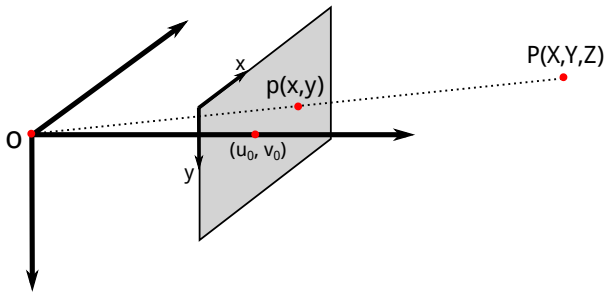

Figura 3: Calibración de la cámara. Adaptada de [Lag11].

En la figura 3, se ha re dibujado el modelo de cámara oscura, en el que se ha agregado un cuadro de referencia posicionado en el centro de proyección. De esta figura y del análisis geométrico anteriormente mencionado, se puede derivar que el punto $P(X, Y, Z)$ será proyectado en el plano de la imagen en $(fX/Z, fY/Z)$. Si queremos trasladar éstas coordenadas a píxeles, necesitamos dividir la posición 2D de la imagen por el ancho del píxel px y su altura py respectivamente. Dividiendo la distancia focal f (en coordenadas del mundo real) por px, obtenemos la distancia focal expresada en píxeles fx horizontales; similarmente $fy = f/py$ es definida como la distancia focal expresada en unidades de píxeles verticales. De esta forma, la ecuación proyectiva resulta como en (4), donde (u_0, v_0) es el punto principal que es sumado al resultado con el objetivo de mover el origen a la esquina superior izquierda de la imagen. Éstas ecuaciones, pueden ser reescritas en forma de matriz a través de la introducción de coordenadas homogéneas, en el que un punto 2D es representado por un vector en R^3 y un punto 3D por uno en R^4, en donde la componente extra es un factor de escala arbitrario.

$$\begin{aligned} x &= \frac{f_x X}{Z} + u_0 \\ y &= \frac{f_y Y}{Z} + v_0 \end{aligned} \tag{4}$$

La ecuación proyectiva reescrita se puede observar en (5).

$$s = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{5}$$

Los **parámetros intrínsecos** de la cámara se encuentra representados en la primer matriz de la expresión (5) y resultan constantes para un sistema lente-camara mientras que la segunda matriz es denominada matriz de proyección. Cuando el cuadro de referencia, no está en el centro de proyección de la cámara, necesitamos agregar una matriz de rotación R de 3x3 y un vector de traslación t de 3x1. Estas dos matrices describen las transformaciones rígidas que deben ser aplicadas a un punto 3D con el objetivo de llevarlo al marco de referencia de la cámara. Luego, se puede reescribir la ecuación de proyección de forma más general como en (6)

$$s = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_1 & r_2 & r_3 & t_1 \\ r_4 & r_5 & r_6 & t_2 \\ r_7 & r_8 & r_9 & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{6}$$

Las componentes de rotación y traslación, son llamados **parámetros extrínsecos** de la calibración y resultan ser diferentes para cada punto de vista.

2.4. Matriz fundamental de un par de imágenes

Al tomar dos imágenes de una misma escena desde diferentes puntos de vista, existe una importante relación proyectiva entre estas y la escena. Estas imágenes, pueden haber sido obtenidas mediante la misma cámara (tomando la fotografía desde dos puntos de vistas diferentes), o mediante dos cámaras posicionadas en diferentes lugares observando el mismo punto.

Si consideramos dos cámaras observando el mismo punto X en una escena, como se puede ver en la Fig. 4, hemos visto que podemos encontrar la proyección del punto en los planos de las imágenes x y x' correspondiente a la posición de la cámara en O y O' respectivamente, trazando una línea de unión entre X y el centro de la cámara.

Ahora, si conocemos la proyección del punto X sobre el plano p, pero desconocemos la ubicación de X y su proyección correspondiente sobre el plano p' como se muestra en la Fig. 5, el punto X puede ser localizado en cualquier lugar de la línea de unión entre O y x. Esto, implica que si queremos buscar la correspondencia de un punto de una imagen dada p en otra p', debemos buscar a lo largo de la linea de proyección l' en el plano de la segunda imagen. Esta línea imaginaria l' recibe el nombre de **línea epipolar** del punto x.

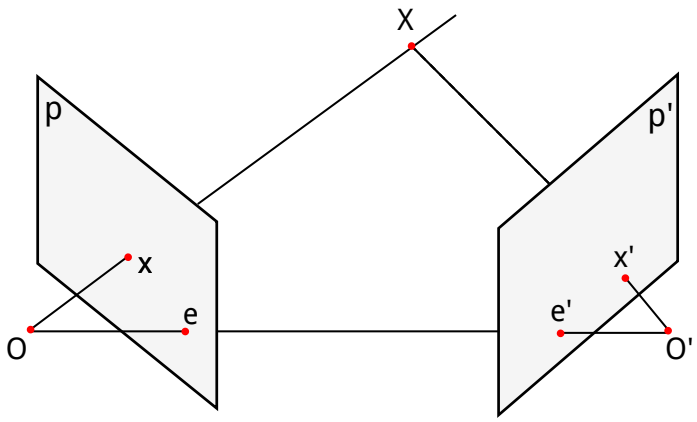

Figura 4: Geometría epipolar: O y O' corresponden a la posición de la cámara en un instante; X representa un punto de la escena a fotografiar y p y p' son el plano imagen obtenido de la proyección del punto X con la cámara en la posición O y O' respectivamente. Adaptada de [Lag11].

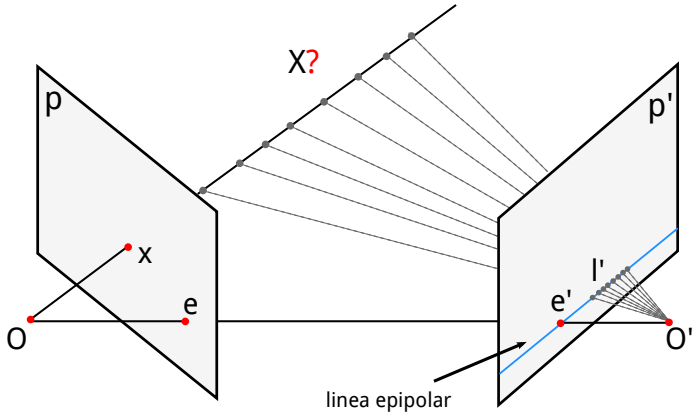

Figura 5: Línea epipolar: Se desconoce la posición del punto X en la escena. Se conoce la proyección de punto X sobre el plano p pero no sobre p'. Adaptada de [Lag11].

El párrafo anterior, define una restricción fundamental que deben cumplir los puntos correspondientes: la correspondencia de un punto dado en una vista, debe estar en la línea epipolar del punto en la otra vista, en donde la orientación exacta de la línea epipolar viene dada por la posición respectiva de las dos cámaras, de esta forma se puede aseverar que la posición de las líneas epipolares, caracterizan la geometría de un sistema con dos puntos de vistas. Además, todas las líneas epipolares pasan a través del mismo punto el cual recibe el nombre de **epipolo**.

Matemáticamente, se puede demostrar [Lag11] que la relación entre un punto de una imagen y su correspondiente línea epipolar, puede ser expre-

sada usando una matriz F de 3x3 como la (7). En geometría proyectiva, una línea 2D viene representada por un vector en R^3 que corresponde al conjunto de puntos 2D (x', y') que satisfacen la ecuación $l'_1x' + l'_2y' + l'_3 = 0$, donde la prima denota que la línea pertenece a la segunda imagen. Consecuentemente, la **matriz fundamental** F, mapea un punto de una imagen 2D de una vista, a una linea epipolar en la otra vista; es decir que esta matriz proporciona para un punto en una imagen, la ecuación de la línea en la que su punto correspondiente (en la otra vista) puede ser hallado. Finalmente, para estimar la matriz fundamental de un par de imágenes se procede mediante la resolución de un sistema de ecuaciones.

$$\begin{bmatrix} l'_1 \\ l'_2 \\ l'_3 \end{bmatrix} = F \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{7}$$

Sea q un punto expresado en coordenadas homogéneas, q' el punto correspondiente a q y F la matriz fundamental entre las dos vistas; dado que q' se encuentra en la línea epipolar Fq, la ecuación (8) recibe el nombre de **restricción epipolar**. Esta ecuación pone en relevancia la relación existente entre dos puntos correspondientes y es mediante la misma que se puede calcular las entradas de la matriz. Dado que las entradas de la matriz F se encuentran afectadas por un factor de escala, sólo hay ocho entradas a estimar (la novena puede ser arbitrariamente establecida a 1). Cada correspondencia se refleja en una ecuación, luego, con ocho correspondencias conocidas la matriz puede ser estimada resolviendo el conjunto de ecuaciones lineales. Pero aún se puede adicionar una restricción más: matemáticamente la matriz F mapea un punto 2D en una línea 1D (es decir líneas que se intersectan en un punto en común). El hecho de que estas lineas epipolares pasan a través de un único punto (el epipolo) impone una restricción a dicha matriz reduciendo así el número de correspondencias requeridas para realizar la estimación a siete ecuaciones y por lo tanto, siete correspondencias.

$$q'^t F q = 0 \tag{8}$$

Debemos mencionar que la selección de un conjunto de correspondencias apropiadas de la imagen, resulta ser de vital importancia para obtener una estimación precisa de la matriz fundamental. En general, las correspondencias deberían estar bien distribuidas a través de las imágenes y deberían incluir puntos a diferentes profundidades de la escena, en caso contrario, la solución no será la esperada resultando en valores espurios.

2.5. Correspondencia de imágenes

A continuación, se expondrá como aprovechar las restricciones epipolares descriptas en la sección 2.4 para obtener una correspondencia de carac-

terísticas más fiables, de tal forma de poder lograr una mejor estimación de la matriz fundamental.

Para lograr esta fiabilidad, sólo se aceptarán que los puntos característicos son coincidentes entre imágenes si los mismos caen sobre la línea epipolar correspondiente a cada uno de ellos. Para comprobar esta condición, nos encontramos con un nuevo problema: en primera instancia, la matriz fundamental debe ser conocida y como sabemos se necesitan buenas coincidencias para estimar bien esta matriz, de esta forma estamos en presencia de un problema mutuamente dependiente. Para poder resolverlo, se propone una solución en el que la matriz fundamental y un conjunto de buenas coincidencias son calculadas de forma conjunta.

A partir de los puntos característicos y sus descriptores, se procede a hallar la correspondencia para cada uno de los puntos ,buscándose las mejores dos coincidencias para cada característica basándose en una métrica de distancia, específicamente la euclídea. Este proceso, se lleva a cabo en ambas direcciones: por cada punto en la primer imagen, se buscan las dos mejores coincidencias en la segunda imagen, y viceversa. Luego, si la distancia medida es muy pequeña para la primer mejor coincidencia y muy grande para la segunda mejor coincidencia, se puede aceptar la primera como la mejor opción. Recíprocamente, si las dos mejores coincidencias son relativamente cercanas en distancia, se desechan ambas mediante un valor de umbral ya que puede haber un error si seleccionamos la primera o la segunda.

Si bien con los pasos descriptos anteriormente se logra reducir de forma considerable la cantidad de coincidencias ambiguas, si se está en presencia de una gran cantidad de buenas coincidencias [Lag11], un significante número de falsas coincidencias aún persiste, para lo cual se realiza un segundo filtrado para poder eliminar ambigüedades o coincidencias no validas como se expone a continuación.

Hasta el momento, se ha obtenido un conjunto de coincidencias de la primer imagen hacia la segunda y otro de la segunda hacia la primera. Así, se extraen las que son válidas en ambos conjuntos mediante un esquema de coincidencias simétrico: un par de coincidencias es aceptada como válida si ambos puntos son la mejor característica de coincidencia en cada una de las imágenes.

Finalmente, se agrega otro paso de filtrado el cual consiste en usar la matriz fundamental con el objetivo de descartar coincidencias que no cumplen con la restricción epipolar. Este test está basado en el método RANSAC (Muestras aleatorias consensuadas, del inglés: Random Sample Consensus) [HZ04, FB81, HZ04, CKY09, Chu05], que puede calcular la matriz fundamental aún cuando valores atípicos están presentes en los conjuntos de datos. Debemos notar que los diferentes pasos para eliminar ambigüedades y coincidencias invalidas, son llevados a cabo con el objetivo de lograr un conjunto de datos fiables que es necesario para lograr una buena estimación de la matriz. En caso de no hacerlo, posiblemente se estaría en presencia de

un resultado no deseado y poco fiable.

2.5.1. RANSAC

El objetivo del algoritmo RANSAC, es estimar una entidad matemática dada desde un conjunto de datos que contiene valores atípicos o incorrectos. Se basa en la idea de seleccionar puntos del conjunto de forma aleatoria llevando a cabo la estimación sólo con este conjunto de datos. El número de puntos seleccionados debe ser el mínimo requerido para estimar la entidad matemática en cuestión, así en el caso de la matriz fundamental, ocho pares de coincidencias resulta ser la cantidad necesaria (podrían ser siete, pero tomar 8 produce un algoritmo lineal que es más rápido de resolver). Una vez que la matriz fundamental es estimada con estas ocho coincidencias aleatorias, todas las demás coincidencias en el conjunto son probadas con la restricción epipolar que se deriva de esta matriz. Todas las coincidencias que cumplen esta restricción, es decir, coincidencias para las cuales la característica correspondiente se encuentra a una distancia cercana de la línea epipolar, son identificadas para formar parte del "conjunto soporte" de la matriz fundamental calculada.

La idea central detrás del algoritmo RANSAC es que cuanto mayor sea el "conjunto soporte", mayor será la probabilidad de que la matriz calculada sea la correcta, de manera que si una (o más) de las coincidencias aleatorias seleccionadas son malas coincidencias, la matriz fundamental calculada seguramente será una mala estimación y por lo tanto, se espera que el conjunto soporte sea menor. Este proceso es repetido varias veces y finalmente se toma como más probable a la matriz con el soporte más grande.

Resumiendo, el objetivo es tomar varias veces ocho coincidencias de forma aleatoria, de manera que eventualmente seleccionemos ocho buenas coincidencias que nos provean de un conjunto soporte grande. Dependiendo de la cantidad de falsas coincidencias que haya en todo el conjunto de datos, la probabilidad de seleccionar un conjunto de ocho coincidencias correctas diferirá. Sin embargo, sabemos que cuanto más selecciones hagamos, mayor será la confianza en encontrar por lo menos un conjunto de buenas coincidencias.

Cuanto mayor sea la cantidad de coincidencias válidas en el conjunto inicial, mayor es la probabilidad de que obtengamos la matriz fundamental correcta con el algoritmo RANSAC, es por esto que antes de buscar la matriz fundamental se aplican todos los pasos anteriormente descriptos de manera de filtrar las posibles coincidencias espurias.

2.6. Homografía

En secciones anteriores, se ha mostrado como calcular la matriz fundamental de un par de imágenes desde un conjunto de puntos coincidentes.

Existe otra entidad matemática que puede ser calculada desde pares de coincidencias llamada Homografía [Lag11, HZ04].

Una homografía, es un mapeo uno a uno entre dos imágenes que queda definida por ocho parámetros y describe la proyección perspectiva entre dos imágenes tomados desde diferente puntos de vista cuando:

- El movimiento de la cámara es de rotación pura y
- La cámara observa una escena plana.

Si consideramos nuevamente la relación proyectiva existente entre un punto 3D y su imagen en la cámara que se ha introducido en 2.2, se ha aprendido que esta relación es expresada por una matriz de $3x4$. Si se considera que las vistas de la escena están separadas puramente mediante una rotación, se puede observar que la cuarta columna de la matriz de parámetros extrínsecos se convierte en 0 (traslación nula). Como resultado, la relación proyectiva en este caso especial da la matriz homografía H de $3x3$.

Si consideramos el conjunto de puntos (x_i, y_i, z_i) en coordenadas homogéneas perteneciente a la primer imagen y sabemos que mapean a un conjunto de puntos (x'_i, y'_i, z'_i) en la segunda imagen, la relación entre las dos imágenes es una homografía si se cumple la ecuación (9)

$$\begin{bmatrix} x'_i \\ y'_i \\ z'_i \end{bmatrix} = H \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} \Longrightarrow \begin{bmatrix} x'_i \\ y'_i \\ z'_i \end{bmatrix} = \begin{bmatrix} h_{11} & h_1 2 & h_1 3 \\ h_{21} & h_2 2 & h_2 3 \\ h_{31} & h_3 2 & h_3 3 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} \tag{9}$$

Así, se puede interpretar de (9) que la Homografia H mapea coordenadas x en una imagen a coordenadas x' en otra. En ambos casos se trata de coordenadas homogéneas –cada punto de la pantalla es tratado como un rayo que pasa a través del centro de la cámara.

Una propiedad característica de la homografía es que siempre mapea una linea recta en otra linea recta, pero el paralelismo no es necesariamente preservado.

A pesar que la matriz H contiene nueve elementos, resulta ambigua al escalarla – el uso de coordenadas homogéneas significa que cualquier múltiplo de la homografía, tendría el mismo efecto. Agregando un factor de escala s, la expresión (9) queda como (10). Consecuentemente, como hay solo ocho elementos independientes, se puede obtener la homografía que relaciona dos imágenes usando solo cuatro correspondencias en donde cada correspondencia genera dos ecuaciones lineales.

$$\begin{bmatrix} sx'_i \\ sy'_i \\ s' \end{bmatrix} = H \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \tag{10}$$

2.6.1. Implementación

Mediante la función *cvFindHomography*[1] provista por OpenCV se procede a buscar la transformación perspectiva entre dos planos. A continuación se presenta el encabezado y sus parámetros:

```
void cvFindHomography(const CvMat* srcPoints,
                      const CvMat* dstPoints,
                      CvMat* H int method=0,
                      double ransacReprojThreshold=0,
                      CvMat* status=NULL
                     )
```

- **param srcPoints:** Coordenadas del punto en el plano original, es un arreglo de 1 canal de $2xN$, $Nx2$, $3xN$ o $Nx3$ (los últimos casos son para la representación en coordenadas homogéneas), donde N es el número de puntos. Un arreglo de 2 o 3 canales de tamaño $1xN$ o $Nx1$ también puede ser pasado.
- **param dstPoints:** Coordenadas del punto en el plano destino. Arreglo de 1 canal de $2xN$, $Nx2$, $3xN$ o $Nx3$, o un arreglo de 2 o 3 canales de tamaño $1xN$ o $Nx1$.
- **param H:** matriz homografía de salida ($3x3$).
- **param method:**
  - **0:** método regular que usa todos los puntos.
  - **CV_RANSAC:** método robusto basado en RANSAC.
  - **CV_LMEDS:** método robusto de la menor mediana.
- **param ransacReprojThreshold** El valor de error de re proyección máximo permitido para tratar a un par de puntos como inliers (usado únicamente con el método RANSAC). Esto es:

  $$||dstPoints_i - convertPointHomogeneous(HsrcPoints_i)|| > ransacReprojThreshold$$

  luego el punto $_i$ es considerado como outlier. Si $srcPoints$ y $dstPoints$ son medidos en píxeles, un valor con sentido usualmente usado está en el rango 1 a 10.
- **param status:** máscara de salida opcional devuelta por el método robusto (**CV_RANSAC** o **CV_LMEDS**). Notar que los valores de la máscara de entrada se han ignorado.

  La función busca la transformación perspectiva H entre los planos fuente y destino:

[1] `http://opencv.willowgarage.com/documentation/camera_calibration_and_3d_`

$$s_i = \begin{bmatrix} x'_i \\ y'_i \\ 1 \end{bmatrix} \sim H \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}$$

por lo que el error de retroproyección es minimizado:

$$\sum_i \left(x'_i - \frac{h_{11}x_i + h_{12}y_i + h_{13}}{h_{31}x_i + h_{32}y_i + h_{33}} \right)^2 + \left(y'_i - \frac{h_{21}x_i + h_{22}y_i + h_{23}}{h_{31}x_i + h_{32}y_i + h_{33}} \right)^2$$

Si el parámetro *method* se establece al valor por defecto 0, la función utiliza todos los pares de puntos para calcular la homografía inicial estimada con un simple esquema de mínimos cuadrados.

Sin embargo, si no todos los pares de puntos $(srcPoints_i, dstPoints_i)$ encajan en la transformación perspectiva (por ejemplo hay algunos valores atípicos), esta estimación inicial será pobre. En este caso, se puede utilizar uno de los dos métodos robustos: $RANSAC$ o $LMeDS$. Estos métodos, prueban muchos subgrupos diferentes de puntos correspondientes (de a 4 pares cada uno) de forma aleatoria, estiman la homografía usando el subconjunto y el algoritmo simple de mínimos cuadrados y luego calculan la calidad de la matriz homográfica calculada. Luego, el mejor subconjunto es usado para producir la estimación inicial de la matriz de homografía y la máscara de valores válidos (inliers) y no válidos (outliers).

El método RANSAC necesita un umbral para distinguir entre posibles valores válidos y atípicos. El método LMeDS no necesita ningún umbral, pero funciona correctamente solo cuando hay más de un 50 % de valores buenos (inliers). Finalmente, si se está seguro que en las características calculadas sólo hay presente algo de ruido pequeño, pero no hay valores atípicos, el métodos por defecto puede ser la mejor opción.

3. Superposición de imagen virtual

Como se explicó anteriormente, una vez que sea logrado obtener la homografía, todos los puntos en una vista pueden ser convertidos a la segunda vista usando la relación (10).

De esta forma, se puede utilizar una imagen que se quiera sobre imponer aplicándole la inversa de la homografía calculada anteriormente de forma de obtener los puntos correspondientes donde debería dibujarse en el flujo de video y así lograr el efecto de sobre imponer el objeto virtual en la realidad. Existe una función provista por la librería OpenCV que aplica una transformación perspectiva a una imagen denominada cvWarpPerspective()[2] y que es la usada para llevar a cabo esta tarea.

reconstruction.html#findhomography

[2] http://opencv.willowgarage.com/documentation/cpp/imgproc_geometric_image_transformations.html#cv-warpperspective

Referencias

[BK08] Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, 1st edition, October 2008.

[BL97] Jeffrey S. Beis and David G. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, CVPR '97, pages 1000–, Washington, DC, USA, 1997. IEEE Computer Society.

[Chu05] Ondřej Chum. Two-view geometry estimation by random sample and consensus. Number CTU–CMP–2005–19, page 92, Prague, Czech Republic, September 2005.

[CKY09] S. Choi, T. Kim, and W. Yu. Performance evaluation of RANSAC family. In *BMVC*, pages xx–yy, 2009.

[FB81] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[FBF77] Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Softw.*, 3(3):209–226, September 1977.

[HZ04] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, June 2004.

[Lag11] Robert Laganière. *OpenCV 2 Computer Vision Application Programming Cookbook*. Packt Publishing, June 2011.

[LMGY04] Ting Liu, Andrew W. Moore, Alexander Gray, and Ke Yang. An investigation of practical approximate nearest neighbor algorithms. pages 825–832. MIT Press, 2004.