Índice general

CAPÍTULO 1

Modelo de cámara y estimación de pose monocular

1.1. Introducción

Se le llama "estimación de pose" al proceso mediente el cual se calcula en qué punto del mundo y con qué orientación se encuentra determinado objeto respecto de un eje de coordenadas previamente definido al que se lo llama *ejes del mundo*. Las aplicaciones de realidad aumentada requieren de un modelado preciso del entorno respecto de estos ejes, para poder ubicar correctamente los agregados virtuales dentro del modelo y luego dibujarlos de forma coherente en la imagen vista por el usuario. El objeto cuya estimación de pose resulta de mayor importancia es la cámara, ya que por ésta es por donde se mira la escena y es respecto de ésta que los objetos virtuales deben ubicarse de manera consistente. Una forma de estimar la pose de la cámara es mediante el uso de las imágenes capturadas por ella misma.
Asimismo, el concepto "monocular" hace referencia al uso de una sola cámara, ya que es posible trabajar con más de una.
Para poder estimar la pose de una cámara, resulta necesario modelarla adecuadamente ya que no todas las cámaras son iguales. El modelo más comunmente utilizado es el denominado *pin-hole*. Para modelar completamente la cámara se deben estimar ciertos *parámetros intrínsecos* a ésta, y eso se logra luego de realizados ciertos experimentos. A la estimación de estos parámetros se le denomina *calibración de la cámara*.

1.2. Calibración de cámara: modelo pin-hole [1]

1.2.1. Fundamentos y definiciones

Este modelo consiste en un centro óptico C, en donde convergen todos los rayos de la proyección y un plano imagen en el cual la imagen es proyectada. Se define "distancia focal" (f) como la distancia entre el centro óptico C y el cruce del eje óptico por el plano imagen (punto P). Ver imagen 1.1.

Para modelar el proceso de proyección (proceso en el que se asocia al punto **M** del mundo, un punto **m** en la imagen), es necesario referirse a varias transformaciones y varios ejes de coordenadas.

- *Coordenadas del mundo*: son las coordenadas que describen la posición 3D del punto **M**. Se definen respecto de los *ejes del mundo* (X_m, Y_m, Z_m). La elección de los ejes del mundo es arbitraria.

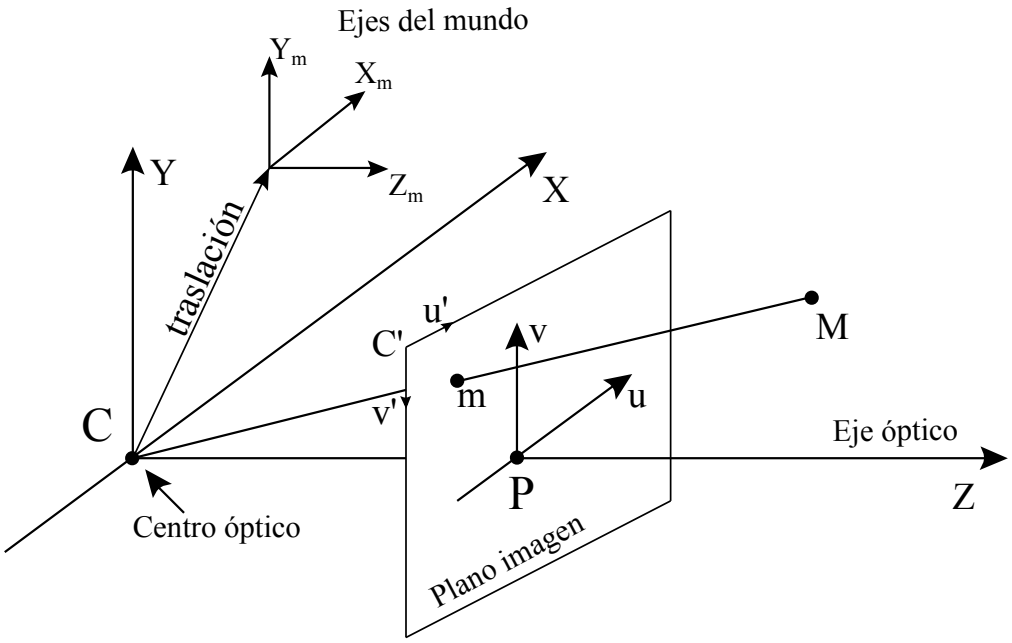

Figura 1.1: Modelo de cámara pin-hole.

- *Coordenadas de la cámara*: son las coordenadas que describen la posición del punto **M** respecto de los ejes de la cámara (X,Y,Z).
- *Coordenadas de la imagen*: son las coordenadas que describen la posición del punto 2D, **m**, respecto del centro del plano imagen, P. Los ejes de este sistema de coordenadas son (u,v).
- *Coordenadas normalizadas de la imagen*: son las coordenadas que describen la posición del punto 2D, **m**, respecto del eje de coordenadas (u',v') situado en la esquina superior izquierda del plano imagen.

La transformación que lleva del punto **M**, expresado respecto de las coordenadas del mundo, al punto **m**, expresado respecto del sistema de coordenadas normalizadas de la imagen, se puede ver como la composición de dos transformaciones menores. La primera, es la que realiza la proyección que transforma a un punto definido respecto del sistema de coordenadas de la cámara (X,Y,Z) en otro punto sobre el plano imagen expresado respecto del sistema de coordenadas normalizadas de la imagen (u',v'). Véase que una vez calculada esta transformación, es una constante característica de cada cámara. Se le llama al conjunto de valores que definen esta transformación *parámetros intrínsecos* de la cámara. La segunda, es la transformación que lleva de expresar a un punto respecto de los ejes del mundo (X_m,Y_m,Z_m), a los ejes de la cámara (X,Y,Z). Esta última transformación varía conforme se mueve la cámara (respecto de los ejes del mundo) y el conjunto de valores que la definen es denominado *parámetros extrínsecos* de la cámara. Del cálculo de estos parámetros es que se obtiene la estimación de la pose de la cámara.
De lo anterior se concluye rápidamente que si se le llama PY a la matriz proyección total, tal que:

$$m = PY.M,$$

entonces:

$$PY = I.E$$

donde I corresponde a la matriz proyección asociada a los parámetros intrínsecos y E corresponde a la matriz asociada a los parámetros extrínsecos. Ambos juegos de parámetros acarrean información

información muy valiosa:

- **Parámetros extrínsecos:** pose de la cámara.
  - Traslación: ubicación del centro óptico de la cámara respecto de los ejes del mundo.
  - Rotación: rotación del sistema de coordenadas de la cámara (X,Y,Z), respecto de los ejes del mundo.
- **Parámetros intrínsecos:** parámetros propios de la cámara. Dependen de su geometría interna y de su óptica.
  - Punto principal (P = $[u'_P, v'_P]$): es el punto intersección entre el eje óptico y el plano imagen. Las coordenadas de este punto vienen dadas en píxeles y son expresadas respecto del sistema normalizado de la imagen.
  - Factores de conversión píxel-milímetros (d_u, d_v): indican el número de píxeles por milímetro que utiliza la cámara en las direcciones u y v respectivamente.
  - Distancia focal (f): distancia entre el centro óptico (**C**) y el punto principa (**P**). Su unidad es el milímetro.
  - Factor de proporción (s): indica la proporción entre las dimensiones horizontal y vertical de un píxel.

1.2.2. Matriz de proyección

En la sección anterior se vió que es posible hallar una "matriz de proyección" PY que dependa tanto de los parámetros intrínsecos de la cámara como se sus parámetros extrínsecos:

$$m = PY.M$$

donde **M** y **m** son los puntos ya definidos y vienen expresados en *coordenadas homogéneas*. Por más información acerca de este tipo de coordenadas ver [8].
Para determinar la forma de la matriz de proyección se estudia cómo se relacionan las coordenadas de **M** con las coordenads de **m**; para hallar esta relación se debe analizar cada transformación, entre los sistemas de coordenadas mencionados con anterioridad, por separado.

- **Proyección 3D - 2D:** de las coordenadas homogéneas del punto **M** expresadas en el sistema de coordenadas de la cámara (X_0, Y_0, Z_0, T_0), a las coordenadas homogéneas del punto **m** expresadas en el sistema de coordenadas de la imagen (u_0, v_0, s_0):
  Se desprende de la imagen 1.1 y algo de geometría proyectiva la siguiente relación entre las coordenadas en cuestión y la distancia focal (f):

$$\frac{f}{Z_0} = \frac{u_0}{X_0} = \frac{v_0}{Y_0}$$

  A partir de la relación anterior:

$$\begin{pmatrix} u_0 \\ v_0 \end{pmatrix} = \frac{f}{Z_0} \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix}$$

Expresado en forma matricial, en coordenadas homogéneas:

$$\begin{pmatrix} u_0 \\ v_0 \\ s_0 \end{pmatrix} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \\ 1 \end{pmatrix}$$

- **Transformación imagen - imagen:** de las coordenadas homogéneas del punto **m** expresadas respecto del sistema de coordenadas de la imagen (u_0, v_0, s_0), a las coordenadas homogéneas de él mismo pero expresadas respecto del sistema de coordenadas normalizadas de la imagen (u'_0, v'_0, s'_0):

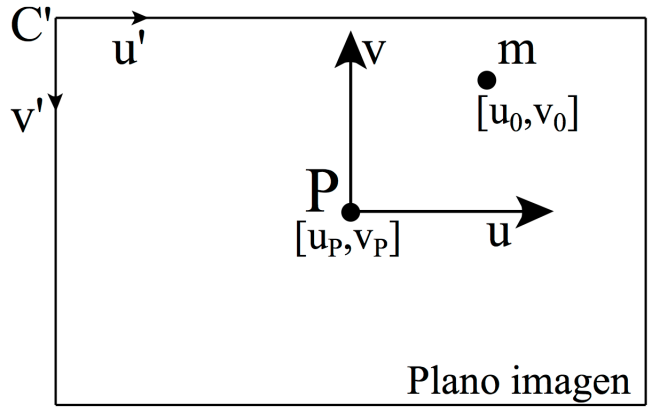

Figura 1.2: Relación entre el sistema de coordenadas de la imagen y el sistema de coordenandas normalizadas de la imagen.

Se les suma, a las coordenadas de **m** respecto del sistema de la imagen, la posición del punto P respecto del sistema normalizado de la imgen (u'_P, v'_P). Las coordenadas de **m** dejan de ser expresadas en milímetros para ser expresadas en píxeles. Aparecen los factores de conversión d_u y d_v:

$$u'_0 = d_u.u_0 + u'_P$$
$$v'_0 = d_v.v_0 + v'_P$$

Se obtiene entonces la siguiente relación matricial, en coordenadas homogéneas:

$$\begin{pmatrix} u'_0 \\ v'_0 \\ s'_0 \end{pmatrix} = \begin{pmatrix} d_u & 0 & u'_P \\ 0 & d_v & v'_P \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_0 \\ v_0 \\ 1 \end{pmatrix}$$

- **Matriz de parámetros intrínsecos** (I)**:** de las coordenadas homogéneas del punto **M** expresadas en el sistema de coordenadas de la cámara (X_0, Y_0, Z_0, T_0), a las coordenadas homogéneas del punto **m** expresadas respecto del sistema de coordenadas normalizadas de la imagen (u'_0, v'_0, s'_0):
  Se obtiene combinando las dos últimas transformaciones. Nótese que como ya se aclaró, depende únicamente de parámetros propios de la construcción de la cámara:

$$I = \begin{pmatrix} d_u.f & 0 & u'_P & 0 \\ 0 & d_v.f & v'_P & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Nota: De forma genérica se puede agregar a la matriz de parámetros intrínsecos del modelo *pin-hole* un parámetro **s** llamado en inglés *skew parameter*, o "parámetro de sesgado" en Español. Este parámetro toma valores distintos de cero muy rara vez, pues modela los casos en los que los ejes x e y de los píxeles de la cámara no son perpendiculares entre sí. En casos realistas, $s \neq 0$ cuando por ejemplo se toma una fotografía de una fotografía. La matriz de parámetros intrínsecos, tomando en cuenta este parámetro, tendrá la forma:

$$I = \begin{pmatrix} d_u.f & s & u'_P & 0 \\ 0 & d_v.f & v'_P & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

- **Matriz de parámetros extrínsecos** (E)**:** de las coordenadas homogéneas del punto **M** expresadas respecto del sistema de coordenadas del mundo $(X_{m0}, Y_{m0}, Z_{m0}, T_{m0})$, a las coordenadas homogéneas de él mismo pero expresadas respecto del sistema de coordenadas de la cámara (X_0, Y_0, Z_0, T_0):
  Se obtiene de estimar la pose de la cámara respecto de los ejes del mundo y es la combinación de, primero una rotación R, y luego una traslación T. Se obtiene entonces la siguiente representación matricial:

$$\begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \\ T_0 \end{pmatrix} = \begin{pmatrix} R & T \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_{m0} \\ Y_{m0} \\ Z_{m0} \\ T_{m0} \end{pmatrix}$$

  donde la matriz de parámetros extrínsecos desarrollada toma la forma:

$$E = \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- **Matriz de proyección** (PY)**:** de las coordenadas homogéneas del punto **M** expresadas respecto del sistema de coordenadas del mundo $(X_{m0}, Y_{m0}, Z_{m0}, T_{m0})$, a las coordenadas homogéneas del punto **m** expresadas respecto del sistema de coordenadas normalizadas de la imagen (u'_0, v'_0, s'_0):
  Es la proyección total y se obtiene combinando las dos transformaciones anteriores:

$$\begin{pmatrix} u'_0 \\ v'_0 \\ s'_0 \end{pmatrix} = \begin{pmatrix} d_u.f & 0 & u'_P & 0 \\ 0 & d_v.f & v'_P & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} . \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} . \begin{pmatrix} X_{m0} \\ Y_{m0} \\ Z_{m0} \\ T_{m0} \end{pmatrix}$$

1.3. Distorsión introducida por las lentes

Hasta el momento se asumió que el modelo lineal presentado para la proyección de cualquier punto del mundo en el plano imagen de la cámara es lo suficientemente preciso en todos los casos. Sin embargo, en casos reales, y cuando las lentes de las cámaras en cuestión no son del todo buenas, la distorsión introducida por estas se hace cada vez más evidente. Dado un

punto **M** de coordenadas (X,Y,Z) respecto de los ejes de la cámara, se le llama distorsión a la diferencia entre la proyección ideal de dicho punto en el plano imagen (u_0, v_0) y su proyección real $(\tilde{u_0}, \tilde{v_0})$. La más común de todas, es la denominada "distorsión radial", ya que su magnitud depende del radio medido desde el punto principal del plano imagen, hasta las coordenadas del punto en cuestión.

La forma de solucionar el presente problema es realizar una corrección de la distorsión, modelando a la misma de la siguiente manera:

$$\begin{pmatrix} \tilde{u_0} \\ \tilde{v_0} \end{pmatrix} = L(r).\begin{pmatrix} u_0 \\ v_0 \end{pmatrix},$$

donde r es la distancia radial $\sqrt{u_0^2+v_0^2}$ y $L(r)$ es un factor de distorsión que depende únicamente del radio r. Si se desarrolla la ecuación anterior, y se expresa en píxeles en la imagen (sistema de coordenadas normalizadas de la imagen), se obtiene lo siguiente:

$$\tilde{u_0}' = u'_P + L(r)(u'_0 - u'_P)$$
$$\tilde{v_0}' = v'_P + L(r)(v'_0 - v'_P)$$

donde $(\tilde{u_0}', \tilde{v_0}')$ son las coordenadas reales de la proyección medidas en píxeles, (u'_0, v'_0) son las coordenadas ideales de la proyección medidas también en píxeles y (u'_P, v'_P) son las coordenadas del punto principal. Véase que en este caso $r = \sqrt{(u'_0 - u'_P)^2 + (v'_0 - v'_P)}$.

La función $L(r)$ es definida sólo para valores positivos de r y $L(0) = 1$. Una aproximación a la función arbitraria $L(r)$ puede ser una expansión de Taylor: $L(r) = 1 + k_1 r + k_2 r^2 + k_3 r^3 + ...$. Finalmente, a la hora de calcular los parámetros intrínsecos de una cámara, también deben ser estimados sus coeficientes de distorsión radial $\{k_1, k_2, k_{3,4}, ...\}$.

1.4. Métodos para la calibración de cámara

Como se vió algunos párrafos atrás, el proceso mediante el cual se calculan los parámetros intrínsecos reales de una cámara es denominado "calibración de cámara". Existen varios métodos para dicho proceso, entre los que se destacan el método de Zhang [10] y el método de Heikkilä y Silvén [12]. Como en este proyecto se utilizó una implementación basada en el primero de ellos (se hablará de dicha implementación un poco más adelante), este será explicado brevemente a continuación.

Bibliografía

[1] J. García Ocón. Autocalibración y sincronización de múltiples cámaras plz. 2007.

[2] B. Furht. *The Handbook of Augmented Reality*. 2011.

[3] C. Avellone and G. Capdehourat. Posicionamiento indoor con señales wifi. 2010.

[4] Philip David, Daniel Dementhon, Ramani Duraiswami, and Hanan Samet. Simultaneous pose and correspondence determination using line features. pages 424–431, 2003.

[5] Philip David, Daniel Dementhon, Ramani Duraiswami, and Hanan Samet. Softposit: Simultaneous pose and correspondence determination. pages 424–431, 2002.

[6] Daniel F. DeMenthon and Larry S. Davis. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15:123–141, 1995.

[7] R. Grompone von Gioi, J. Jakubowicz, J. M. Morel, and G. Randall. Lsd: A fast line segment detector with a false detection control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):722–732, April 2010.

[8] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[9] V. Lepetit and P. Fua. Monocular model-based 3d tracking of rigid objects: A survey. *Foundations and Trends in Computer Graphics and Vision*, 1(1):1–89, 2005.

[10] Zhengyou Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *ICCV*, pages 666–673, 1999.

[11] Denis Oberkampf, Daniel F. DeMenthon, and Larry S. Davis. Iterative pose estimation using coplanar feature points. *Comput. Vis. Image Underst.*, 63(3):495–511, may 1996.

[12] Jane Heikkilä and Olli Silvén. A four-step camera calibration procedure with implicit image correction. In *1997 Conference on Computer Vision and Pattern Recognition (CVPR 97), June 17-19, 1997, San Juan, Puerto Rico*, page 1106. IEEE Computer Society, 1997.