Capitolo 2: Il Modello Pinhole

Il modello pinhole idealizza la camera come un dispositivo privo di aberrazioni ottiche: ogni punto dello spazio è “proiettato” linearmente sul piano immagine attraverso un unico centro di proiezione. Questa semplicità matematica è alla base di quasi tutti gli algoritmi di visione artificiale e calibrazione fotografica.

Tale modello ha origini antichissime:

- Prima documentazione (V sec. a.C.)
  Il filosofo cinese Mozi (470–390 a.C.) fu tra i primi a descrivere in termini teorici il fenomeno di un'apertura che proietta su una parete oscura un’immagine capovolta della scena esterna.
- Osservazioni in Grecia (IV sec. a.C.)
  Aristotele (384–322 a.C.) menzionò l’effetto della “camera oscura” durante l’osservazione delle eclissi solari: i raggi di luce che filtravano tra le foglie degli alberi formavano sul terreno immagini capovolte del Sole.
- Prima trattazione scientifica (XI sec. d.C.)
  Il matematico e fisico arabo Alhazen realizzò, tra il 1012 e il 1021, esperimenti sistematici con la “stanza buia”, fornendo la prima descrizione geometrica e quantitativa del dispositivo che chiamiamo oggi “pinhole camera”.
- Modello pinhole in chiave moderna
  Il modello pinhole, pur avendo origini antiche, entra ufficialmente in uso nella visione artificiale a partire dai primi anni ’70, quando viene adottato per la calibrazione fotogrammetrica tramite il metodo del Direct Linear Transformation. Il lavoro fondativo è quello di Abdel-Aziz & Karara (1971), che propongono la DLT per mappare coordinate 3D in coordinate di immagine in close-range photogrammetry.
  Nel 1987 Roger Y. Tsai pubblica “A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses”, in cui utilizza esplicitamente il modello pinhole per calibrare camere montate su robot e per la guida di veicoli automatici, quello di Tsai è considerato il primo metodo sistematico di calibrazione “robot-oriented” basato sul modello pinhole, ed è tuttora alla base di gran parte delle tecniche di visione robotica.

2.1: Descrizione del Modello

Il modello si basa su due elementi geometrici fondamentali: un centro di proiezione, ossia un foro infinitamente piccolo attraverso cui passano tutti i raggi luminosi, e un piano immagine posizionato a distanza focale f da quel centro. In questa configurazione, ogni punto $P = (x_1, x_2, x_3)$ nello spazio si proietta in un punto $Q = (y_1, y_2)$ del piano immagine tramite un'unica retta che collega P al foro.

2.2: Equazioni principali del Modello

Dal punto di vista geometrico, si considera un sistema di riferimento 3D con origine nel foro e assi ortogonali (X_1, X_2, X_3) con rispettivamente: "X_1" coordinata del punto lungo l'asse orizzontale della camera, "X_2" coordinata del punto lungo l'asse verticale della camera e "X_3" è detto asse ottico e punta verso la scena.

Il piano immagine è parallelo ad X_1X_2 e interseca l'asse ottico in corrispondenza del punto "$-f$" (oppure, una formulazione equivalente più pratica, in "$+f$", se si introduce un'immagine virtuale).

La distanza focale "f" rappresenta il parametro intrinseco che scala le coordinate normalizzate.

Matematicamente, si ottiene la proiezione prospettica attraverso la similarità di triangoli, tramite la figura vista lungo l'asse X_2 si ricava:

$$y_1 = -\frac{f x_1}{x_3} \qquad\qquad y_2 = -\frac{f x_2}{x_3}$$

Tramite la quale, eliminando la rotazione di 180° intrinseca al modello reale, diventa:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \frac{f}{x_3} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Questa relazione descrive in modo compatto come la profondità x_3 "comprime" le coordinate orizzontali e verticali in funzione di f.

Nel modello pinhole "elementare" abbiamo appena visto la proiezione di un punto $P = (X, Y, Z)$, coordinate non omogenee, sul piano immagine.

Per poter includere in un'unica espressione sia la **rotazione "**R" che la **traslazione "**t" della camera rispetto a un sistema di riferimento globale, è molto comodo "allargare" le nostre coordinate aggiungendo una componente in più, pari a 1.

Possiamo definire quindi: vettore delle coordinate omogenee di un punto nello spazio 3D (*X*) e vettore delle coordinate omogenee di un punto immagine (*x*):

$$X = \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \qquad x = \begin{pmatrix} u \\ v \\ 1 \end{pmatrix}$$

Nel caso del vettore delle coordinate omogenee di un punto nello spazio 3D:

- $(X \quad Y \quad Z)$ sono le coordinate cartesiane del punto nel sistema di riferimento globale.
- La componente "1" consente di rappresentare traslazioni e proiezioni mediante moltiplicazioni matriciali uniformi, esattamente come avviene passando al sistema di coordinate della camera tramite $[R|t]$.

Nel caso del vettore delle coordinate omogenee di un punto immagine:

- $(u \quad v)$ sono le coordinate del punto sul sensore (solitamente espresse in pixel).
- La componente "1" serve a permettere le trasformazioni proiettive (moltiplicazioni matriciali) in modo uniforme.

Supponiamo ora di avere un punto nel "mondo" (ad esempio nel sistema di riferimento di un banco di lavoro) con coordinate: (X, Y, Z).

Immaginiamo ora che la camera, invece, si trova in una certa posizione e con una certa orientazione rispetto a quel sistema.

Sappiamo che:

- R è una matrice 3×3 che ruota il punto dal sistema mondo al sistema camera.
- t è un vettore 3×1 che trasla il punto per tener conto dello spostamento dell'origine.

Abbiamo dunque una matrice $[R|t]$ di dimensione 3×4:

$$[R|t] = \begin{pmatrix} r_{11} & r_{12} & r_{13} \; t_x \\ r_{21} & r_{22} & r_{23} \; t_y \\ r_{31} & r_{32} & r_{33} \; t_z \end{pmatrix}$$

La quale moltiplicandola per "X" otteniamo le coordinate del punto espresso nel sistema di riferimento della camera:

$$\boldsymbol{CoordCam} = \begin{pmatrix} X_c \\ Y_c \\ Z_c \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \; t_x \\ r_{21} & r_{22} & r_{23} \; t_y \\ r_{31} & r_{32} & r_{33} \; t_z \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}$$

Equivalente a scrivere in maniera esplicita il sistema:

$$\begin{cases} X_c = r_{11}X + r_{12}Y + r_{13}Z + t_x \\ Y_c = r_{21}X + r_{22}Y + r_{23}Z + t_y \\ Z_c = r_{31}X + r_{32}Y + r_{33}Z + t_z \end{cases}$$

A questo punto, individuate $(X_c \quad Y_c \quad Z_c)$ nel sistema della camera, vogliamo proiettarlo sul piano immagine tenendo conto però di alcuni fattori:

- f: la lunghezza focale, che scala le coordinate normalizzate in unità di pixel sul sensore.
- (c_x, c_y): il centro dell'immagine (principal point), ossia lo spostamento del punto ottico rispetto al vertice superiore sinistro del sensore.

Per fare ciò abbiamo bisogno di utilizzare la "**matrice dei parametri intrinseci**" (K) la quale ci permette di tradurre le coordinate metriche sul sensore, nelle coordinate discrete in pixel:

$$K = \begin{pmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix}$$

Dove abbiamo che:

- f_x, f_y: sono le lunghezze focali espresse in pixel lungo gli assi orizzontale e verticale.
- s: è il parametro di skew, che misura l'eventuale non ortogonalità tra riga e colonna del sensore.
- c_x, c_y: è il principal point, ossia le coordinate (in pixel) del punto in cui l'asse ottico incide sul sensore. Spesso questo punto non coincide con il centro geometrico del sensore, quindi, le coordinate, tengono conto di un'eventuale traslazione dell'origine del sistema di coordinate dell'immagine.

Moltiplicando K per il vettore omogeneo: $\left(\frac{X_c}{Z_c}, \frac{Y_c}{Z_c}, 1\right)$, otteniamo le coordinate in pixel $(u, v, 1)$.

Riassumendo quindi il tutto in un'unica espressione, chiamata: **“equazione di proiezione della camera”**, si ha che:

$$x \backsim K[R|t]X$$

Dove:

- x: indica le coordinate del punto proiettato sul piano immagine, espresse in forma omogenea.
- K: matrice dei parametri intrinseci, che scala e trasla le coordinate normalizzate in coordinate pixel.
- $[R|t]$: matrice estrinseca, che Ruota e Trasla le coordinate dal sistema mondo al sistema camera.
- X: vettore omogeneo $(X, Y, Z, 1)^T$ del punto nel sistema di riferimento globale.
- $\sim$: indica che, a valle della moltiplicazione, dobbiamo dividere per la terza componente (**normalizzazione omogenea**) per tornare a coordinate 2D effettive.

$$(u, v) = \left(\frac{x}{w}, \frac{y}{w}\right) \qquad se\ (x, y, w)^T = K[R|t]X$$

  - Con w: terza componente del vettore omogeneo risultante dalla proiezione, ed è proprio il fattore di scala che equivale alla profondità del punto rispetto alla camera.

Con questo formalismo, si ha un’unica matrice 3×4 che incapsula posizione, orientazione e proiezione sul sensore: è il cuore di qualsiasi algoritmo di visione artificiale che lavora in contesti reali o robotici.

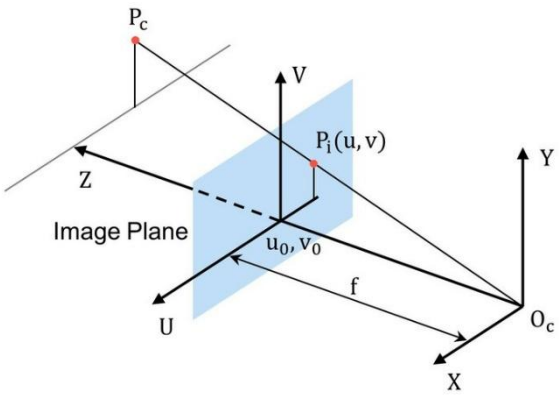

Figura: schema vettoriale del modello pinhole, con centro di proiezione O_c piano immagine a distanza focale f coordinate immagine (u_0, v_0) e proiezione sul punto P_c

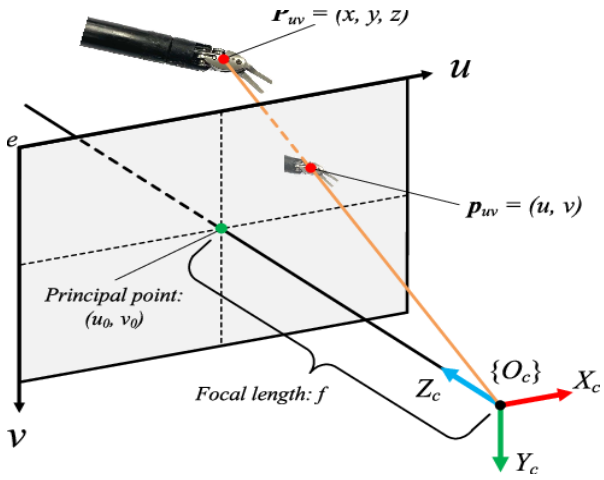

Figura: implementazione reale su un braccio robotico con telecamera, che mostra la proiezione del punto $P_{uv} = (x, y, z)$ nel piano immagine in (u, v) attraverso il foro stenopeico.

2.3: Esempio di un punto 3D proiettato nel piano immagine

Per dare un'idea più chiara del concetto espresso fino ad ora in maniera teorica, andiamo a simulare il meccanismo a livello di calcoli reali.

Dati:

- Scegliamo un punto 3D nel sistema mondo:

$$\boldsymbol{P} = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix} \rightarrow distanza\ in\ metri$$

- Come parametri estrinseci (scelta semplificata per l'esempio), prendiamo:

$$\boldsymbol{R} = \boldsymbol{I_3} \ \rightarrow \ I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$e\ prendiamo\ \boldsymbol{t} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}$$

- Come parametri intrinseci invece usiamo:

$$\boldsymbol{f} = 800\ ed\ anche\ \left(\boldsymbol{c_x}, \boldsymbol{c_y}\right) = (320{,}240)\ \rightarrow distanse\ in\ pixel$$

***Nota**: grazie a tale scelta abbiamo che $f_x = f_y = f$ ed abbiamo anche che Skew è $s = 0$.*

Per procedere, bisogna in primis fare una rototraslazione nel sistema camera:

$$\begin{pmatrix} X_c \\ Y_c \\ Z_c \end{pmatrix} = [R|t] \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} = I_3 \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix}$$

A questo punto troviamo le coordinate normalizzate:

$$\boldsymbol{x} = \frac{X_c}{Z_c} = \frac{1}{5} = 0{,}2$$

$$\boldsymbol{y} = \frac{Y_c}{Z_c} = \frac{2}{5} = 0{,}4$$

A questo punto possiamo passare sul piano immagine:

$$\boldsymbol{u} = fx + c_x\ , \boldsymbol{v} = fy + c_y$$

Quindi:

$$\boldsymbol{u} = 800 \cdot 0{,}2 + 320 = \boldsymbol{480}\ , \boldsymbol{v} = 800 \cdot 0{,}4 + 240 = \boldsymbol{560}$$

Abbiamo dunque concluso che, il punto $P = (1,2,5)$, mappa sul sensore alle coordinate:

$$\boldsymbol{p} = (\boldsymbol{u}, \boldsymbol{v}) = (480,560) \rightarrow in\ pixel$$

Vedendo così in maniera pratica tutti i passaggi: dalla definizione del punto 3D e dei parametri camera, fino al calcolo delle coordinate immagine in pixel.

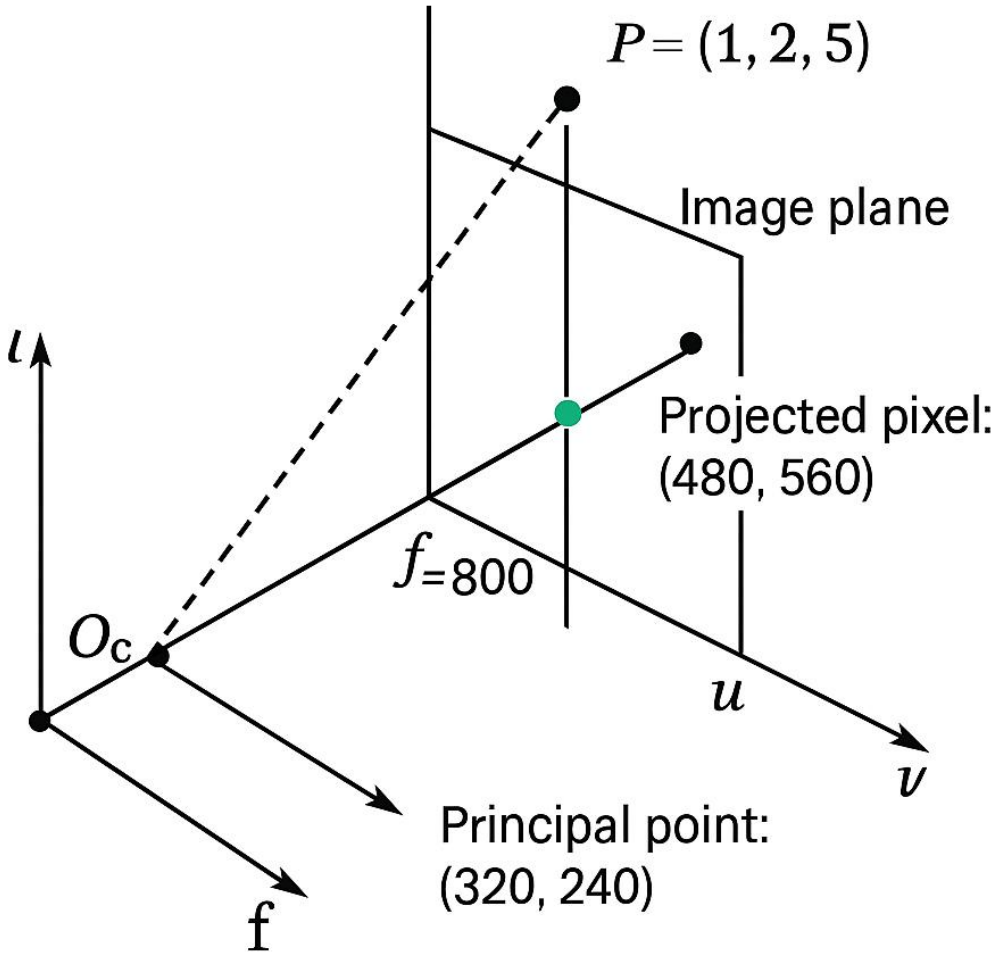

Figura: Diagramma vettoriale dell'esempio, il punto 3D → $P = (1,2,5)$ nel sistema camera $\{O_c\}$ si proietta sul piano immagine, che ha distanza focale $f = 800$, nel pixel di coordinate → $p = (480,560)$.