Appendice A

Il modello di proiezione della videocamera

Per controllare il robot usando le informazioni provenienti dal sistema di visione è necessario comprendere la geometria che descrive il processo di acquisizione dell'immagine[1]. Ogni videocamera contiene una lente che forma una proiezione 2D della scena sul piano immagine dove è posto il sensore (in genere CCD). Questa proiezione comporta una perdita dell'informazione sulla profondità della scena perchè un punto sul piano immagine corrisponde ad un intera retta nello spazio 3D. Sono quindi necessarie altre informazioni per ricavare completamente le caratteristiche della scena 3D: esse possono essere ottenute da videocamere multiple, osservazioni diverse prese da una singola videocamera oppure dalla conoscenza di relazioni geometriche tra i punti del target.

In questa tesi viene preso in considerazione esclusivamente l'utilizzo di una singola videocamera (*single perspective camera*), percui descriveremo accuratamente la geometria che ne regola il comportamento. Tralasciamo gli aspetti riguardanti l'acquisizione contemporanea da diverse videocamere. Con il termine videocamera (*camera*) intenderemo, piuttosto che l'oggetto, la trasformazione che l'oggetto effettua: una videocamera, cioè, trasforma punti del mondo 3D (*object space*) in punti su un'immagine 2D. Il principale tipo di videocamera di nostro interesse è la proiezione centrale (*central projection*).

I modelli di videocamera (*camera models*) che prenderemo in considerazione sono matrici con particolari proprietà che rappresentano il mapping effettuato. Si dividono in due classi principali:

- modelli con centro finito;
- modelli con centro all'infinito;

I modelli che svilupperemo in questo capitolo rappresentano principalmente sensori tipo CCD, ma sono applicabili anche ad altri tipi di vidocamere, come le immagini a raggi-x, i negativi fotografici scansionati etc...

A.1 Modello *basic pinhole*

Consideriamo la proiezione centrale di un punto nello spazio su un piano. Indichiamo come centro di proiezione l'origine di un sistema di coordinate euclidee, e consideriamo il piano $z = f$ chiamato piano immagine (*image plane* o *focal plane*). Nel modello pinhole un punto nello spazio con coordinate $X = (x, y, z)^T$ è trasformato nel punto sul piano immagine dove la linea che collega il punto X al centro di proiezione interseca il piano immagine (vedi fig.A.1). Per triangoli simili, si può facilmente calcolare che il punto

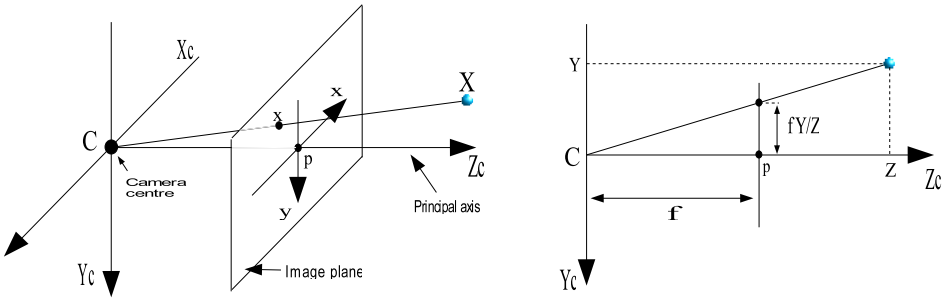

Figura A.1: Geometria della videocamera: C è il centro della videocamera e p è il punto principale. Il centro della videocamera in questo caso coincide con l'origine delle coordinate.

$(x, y, z)^T$ è mappato sul punto $(f\frac{x}{z}, f\frac{y}{z})^T$ sul piano immagine, cioè:

$$\Re^3 \mapsto \Re^2 : \begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} f\frac{x}{z} \\ f\frac{y}{z} \end{pmatrix} \tag{A.1}$$

La A.1 descrive la funzione proiezione centrale dal mondo alle coordinate sull'immagine.

Il centro di proiezione è detto centro della videocamera (*camera centre* o *optical centre*). La linea che parte dal centro della videocamera ed è perpendicolare al piano immagine è detta asse principale (o raggio principale)(*principal axis*) della videocamera. Il punto dove l'asse principale interseca il piano immagine è detto punto principale (*principal point*). Il piano che

attraversa il centro della videocamera, parallelo al piano immagine, è detto piano principale (*principal plane*) della videocamera.

A.1.1 Proiezione centrale in coordinate omogenee

Usando vettori omogenei per rappresentare i punti del mondo e del piano immagine, la proiezione centrale è espressa semplicemente da una funzione lineare tra le loro coordinate omogenee. La A.1 può essere espressa in termini di moltiplicazione tra matrici:

$$\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} f\frac{x}{z} \\ f\frac{y}{z} \\ z \end{pmatrix} = \begin{bmatrix} f & & & 0 \\ & f & & 0 \\ & & 1 & 0 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{A.2}$$

Scomponiamo la matrice come: $diag(f, f, 1)[I|\mathbf{0}]$.

Sia $\mathbf{X}$ il punto nel mondo rappresentato da un vettore omogeneo di 4 elementi $(x, y, z, 1)^T$, $\mathbf{x}$ il punto sull'immagine rappresentato da un vettore omogeneo di 3 elementi e P una matrice omogenea $3x4$ (*camera projection matrix*). Possiamo riscrivere la A.2 come:

$$\mathbf{x} = P\mathbf{X} \tag{A.3}$$

in cui la matrice del modello *basic pinhole* della proiezione centrale è:

$$P = diag(f, f, 1)[I|\mathbf{0}] \tag{A.4}$$

A.1.2 Offset del punto principale

Nell'espressione (A.1) assumiamo che l'origine delle coordinate del piano immagine coincida con il punto principale. In generale questo non è vero (vedi fig.A.2). In generale vale la relazione:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} f\frac{x}{z} + p_x \\ f\frac{y}{z} + p_y \end{pmatrix} \tag{A.5}$$

dove $(p_x, p_y)^T$ sono le coordinate del punto principale. In coordinate omogenee si scrive:

$$\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} fx + zp_x \\ fy + zp_y \\ z \end{pmatrix} = \begin{bmatrix} f & & p_x & 0 \\ & f & p_y & 0 \\ & & 1 & 0 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{A.6}$$

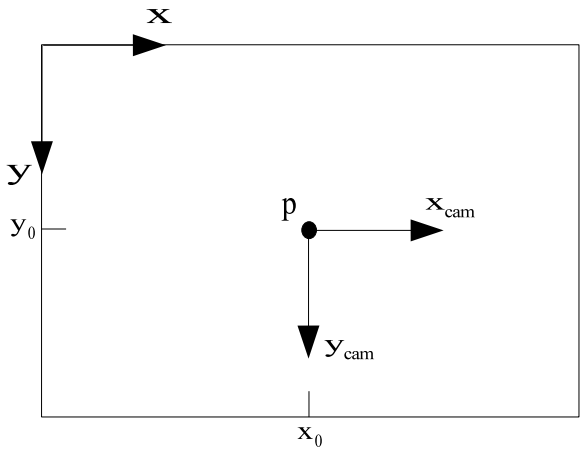

Figura A.2: Coordinate dell'immagine e della terna telecamera.

Indicando con:

$$K = \begin{bmatrix} f & & p_x & 0 \\ & f & p_y & 0 \\ & & 1 & 0 \end{bmatrix} \tag{A.7}$$

Possiamo scrivere in forma concisa:

$$\mathbf{x} = K\,[I|\mathbf{0}]\,\mathbf{X}_{cam} \tag{A.8}$$

K è detta matrice di calibrazione della videocamera (*camera calibration matrix*). Abbiamo scritto $\mathbf{X}_{cam} = (x, y, z, 1)^T$ per enfatizzare che assumiamo che la videocamera sia posta all'origine del sistema di coordinate euclidee con l'asse principale diretto lungo l'asse z. Tale sistema di coordinate può essere detto terna solidale alla videocamera*camera coordinate frame*, che indicheremo con la scrittura $< C >$.

A.1.3 Rotazioni e traslazione della videocamera

In generale i punti nello spazio sono espressi in una terna, diversa da quella solidale alla videocamera, detta terna del mondo (*world coordinate frame*), che indicheremo con $< W >$. I due sistemi sono legati da una rototraslazione (vedi fig.A.3) Sia $\tilde{\mathbf{X}}$ un vettore non omogeneo di 3 coordinate che rappresenta un punto nella terna del mondo, e sia $\tilde{X}_{cam}$ lo stesso punto nella terna solidale alla videocamera. Quindi possiamo scrivere che $\tilde{\mathbf{X}}_{cam} = R(\tilde{\mathbf{X}} - \tilde{\mathbf{C}})$, dove $\tilde{\mathbf{C}}$ rappresenta le coordinate del centro della telecamera rispetto alla terna del mondo, ed R è una matrice di rotazione 3x3 che rappresenta l'orientazione

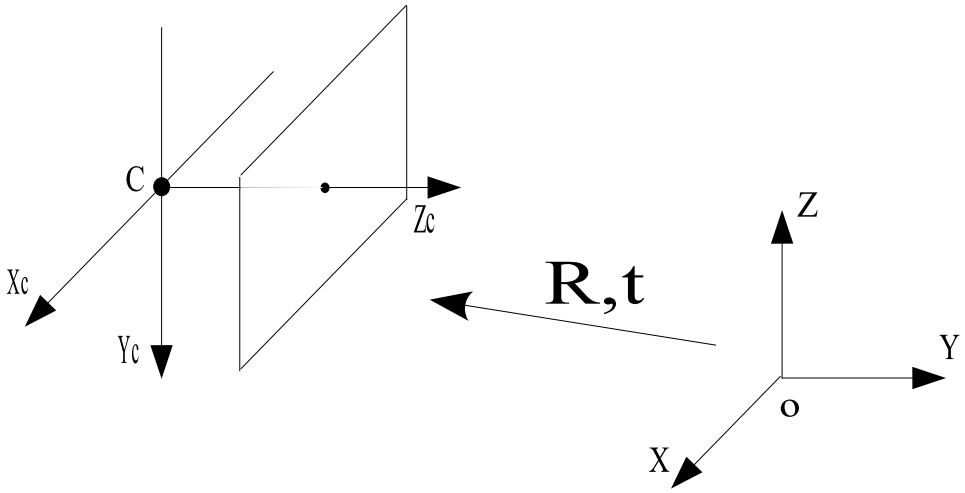

Figura A.3: Trasformazione euclidea tra la terna del mondo e la terna della telecamera.

della terna solidale alla videocamera. Questa equazione può essere scritta in coordinate omogenee come:

$$X_{cam} = \begin{bmatrix} R & -R\tilde{\mathbf{C}} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = \begin{bmatrix} R & -R\tilde{\mathbf{C}} \\ 0 & 1 \end{bmatrix} X \tag{A.9}$$

Sostituendo questa nella (A.8) si ottiene:

$$x = KR\left[I| - \tilde{\mathbf{C}}\right] X \tag{A.10}$$

dove X è ora espresso in terna <W>. Questa è la funzione mappa generale del modello *pinhole*.

Dalle espressioni ricavate si vede che, in generale, una videocamera *pinhole*, $P = KR\left[I| - \tilde{\mathbf{C}}\right]$, ha 9 gradi di libertà:

- 3 per K (f, p_x, p_y);
- 3 per R;
- 3 per $\tilde{\mathbf{C}}$.

I parametri contenuti in K sono detti parametri interni della videocamera (o orientazione interna della videocamera). I parametri di R e $\tilde{\mathbf{C}}$, che legano

la posizione e l'orientamento della videocamera rispetto alla terna $< W >$, sono chiamati parametri esterni (o orientazione esterna).

E' spesso conveniente non indicare esplicitamente il centro della videocamera ma indicare la trasformazione come: $\tilde{X}_{cam} = R\tilde{X} + t$. In questo caso la matrice della videocamera è semplicemente:

$$P = K[R|\mathbf{t}] \tag{A.11}$$

dove, dalla (A.10), $t = -R\tilde{\mathbf{C}}$.

A.1.4 CCD (charge coupled device)

Nel modello *pinhole* che abbiamo analizzato si assume che le coordinate sull'immagine siano euclidee scalate ugualmente su entrambi gli assi. In generale nei CCD, in cui i pixel non sono quadrati, questa ipotesi non è verificata. Quindi, se le coordinate dell'immagine sono misurate in pixel, si introduce un fattore di scala diverso in ogni direzione. In particolare, se il numero di pixel per unità di distanza nelle coordinate dell'immagine è m_x per la direzione x e m_y per la direzione y, allora la trasformazione dalla terna del mondo alle coordinate in pixel si ottiene moltiplicando la (A.7) a sinistra per un fattore extra $diag(x_x, m_y, 1)$. Quindi la matrice di calibrazione per una videocamera CCD ha la forma seguente:

$$K = \begin{bmatrix} \alpha_x & & x_0 \\ & \alpha_y & y_0 \\ & & 1 \end{bmatrix} \tag{A.12}$$

dove $\alpha_x = fm_x$ e $\alpha_y = fm_y$ rappresentano la lunghezza della focale della videocamera nelle due direzioni in termini di pixel. $\tilde{\mathbf{x}}_0 = (x_0, y_0)$ è il punto principale nelle stesse dimensioni, con coordinate $x_0 = m_x p_x$ e $y_0 = m_y p_y$. Una videocamera CCD ha 10 gradi di libertà.

A.2 Modello *scaled orthographic projection*

La *perspective projection* è una relazione non lineare dalle coordinate cartesiane alle coordinate sull'immagine. In alcuni casi è possibile approssimarla con una relazione lineare chiamata *scaled orthographic projection*. Usando questo modello le coordinate immagine di un punto sono date da:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto s \begin{pmatrix} x \\ y \end{pmatrix} \tag{A.13}$$

dove s è un dato fattore di scala. Questo modello è valido nei casi in cui la profondità relativa dei punti nella scena è piccola in confronto alla distanza della videocamera dalla scena, ad esempio un aeroplano in volo sulla terra oppure una videocamera con una lunga focale piazzata a diversi metri dallo spazio di lavoro.

A.3 Modello *affine projection*

Questo modello rappresenta un altro tipo di approssimazione lineare per la *perspective projection*. In questo caso si ha:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto A \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \mathbf{c} \tag{A.14}$$

dove A è una matrice arbitraria 2x3 e $\mathbf{c}$ è un vettore arbitrario di 2 elementi. La *scaled orthographic projection* è una particolare *affine projection*. La proiezione affine non corrisponde a nessuna situazione specifica, ma ha il principale vantaggio di essere un modello lineare non vincolato. Dato un insieme di corrispondenze coordinate nel mondo-coordinate sul piano immagine, A e $\mathbf{c}$ sono calcolate facilmente mediante tecniche di regressione lineare. Quindi il problema della calibrazione è molto semplice per questo modello.

A.4 *Image features* e relativo spazio dei parametri

In computer vision una *image feature* è qualunque caratteristica strutturale della scena che può essere estratta dall'immagine (ad esempio uno spigolo od un contorno). Tipicamente una image feature individuata sul piano immagine corrisponde ad una caratteristica fisica dello stesso oggetto proiettata sul piano immagine.

Definiamo come *image feature parameter* una qualsiasi quantità che può essere calcolata dalla conoscenza di una o più feature, ad esempio relazioni tra regioni e vertici tipo l'area dei poligoni individuati. Nel visual servoing sono in genere utilizzati le coordinate dei punti sul piano immagine, la distanza tra due punti sul piano immagine e l'orientazione della linea che li congiunge, l'area di una superficie proiettata. Per altri parametri vedi [1].

In questo lavoro di tesi restringiamo la nostra attenzione a punti sulla scena (che chiameremo genericamente *feature*) i quali parametri sono le

coordinate sul piano immagine. Una buona *feature* è un punto che può essere localizzato senza ambiguità in differenti osservazioni della scena.

Allo scopo di eseguire *visual servo control* dobbiamo selezionare un insieme di n feature sul piano immagine.