SIFT E GEOMETRIA MULTI-IMMAGINE PER LA MODELLAZIONE TRIDIMENSIONALE AUTOMATICA DA SEQUENZE DI IMMAGINI

Luigi BARAZZETTI, Marco SCAIONI

Politecnico di Milano, Dip. I.I.A.R., P.za Leonardo da Vinci 32, 20133 Milano
luigi.barazzetti@mail.polimi.it , marco.scaioni@polimi.it

Riassunto
L'utilizzo di descrittori capaci di individuare sulle immagini particolari invarianti per rotazione e scala, unitamente alle tecniche di geometria proiettiva, consentono la stima della geometria di presa e la ricostruzione dell'oggetto (Structure & Motion – S&M).
Scopo di questo lavoro è la creazione di un algoritmo in grado di eseguire la Structure and Motion in modalità completamente automatica, a partire da una sequenza di immagini (almeno 2) acquisita mediante una camera digitale calibrata in modo tale che ogni immagine formi una coppia stereo caratterizzata da un elevato ricoprimento con la successiva.
L'algoritmo provvederà in sequenza alla ricerca di punti omologhi mediante l'algoritmo SIFT e alla stima dell'orientamento relativo tre le coppie stereo attraverso la stima robusta della matrice fondamentale per la rimozione di eventuali valori anomali. Successivamente, noti i parametri di calibrazione della camera, è possibile rimuovere l'ambiguità proiettiva per una coppia di immagini scelta come riferimento. Nell'ultima fase, ogni immagine viene concatenata alla precedente mediante una progressiva *space resection* alternata a triangolazione, consentendo così l'aggiunta delle restanti immagini della sequenza.

Abstract
The use of descriptors which are capable of finding scale invariant features on the images and projective geometry techniques allow the estimation of the image orientation and the object reconstruction (Structure and Motion – S&M).
The goal of this work is the creation of an algorithm which is able to perform the Structure and Motion in an automatic way starting from a sequence of images (at least 2) acquired with a digital calibrated camera. It is important that each image forms a stereo pair with the next one.
The algorithm will find homologous points on all images by using the SIFT descriptor, it will estimate the relative orientation parameters between each stereo pair with the robust estimation of the fundamental matrix to perform an outlier removal. Thus, it will provide the 3D coordinates for a stereo pair used as reference with only a scale ambiguity by using the camera calibration parameters.
Lastly the algorithm adds each image by using a progressive concatenation starting from the reference calibrated image pair.

Introduzione

L'analisi delle sequenze di immagini per la ricostruzione 3D degli oggetti è attualmente un obiettivo di primaria importanza sia in fotogrammetria che in *Computer Vision* (Pollefeys *et al.*, 2000). Seppure in entrambe le discipline l'obiettivo finale sia la realizzazione di un modello tridimensionale dell'oggetto, spesso le metodologie di analisi delle immagini e le procedure di orientamento delle stesse sono differenti.
Se in fotogrammetria in genere vengono impiegate le *equazioni di collinearità*, risolte per mezzo di linearizzazione e per via iterativa (quindi tramite l'impiego di parametri approssimati) in *computer*

vision sono state sviluppate da anni tecniche in grado di semplificare il problema rendendolo già lineare in tutte le varie fasi di orientamento e di ricostruzione dell'oggetto. L'utilizzo di questi metodi permette di risolvere i classici problemi fotogrammetrici (orientamento delle immagini, restituzione, ecc.) mediante la scrittura di sistemi lineari, che non richiedono dunque la conoscenza a priori di una soluzione approssimata. Tuttavia, questi non tengono in considerazione alcune problematiche che in fotogrammetria sono ormai ben note e risolte, le quali portano spesso ad errori nella stima della soluzione.
E' necessario specificare che le precisioni richieste nelle due discipline sono in genere molto differenti. Infatti, se in fotogrammetria è lecito aspettarsi precisioni dell'ordine di 10^{-5}, in CV spesso l'obiettivo è la formazione di un modello qualitativo adatto a scopi di visualizzazione (per esempio in una classica applicazione di robotica si ritiene che una precisione del 5% sulla dimensione dell'oggetto sia sufficiente per consentire ad una macchina di evitare gli ostacoli che si possono presentare dinanzi).
Poiché l'elaborazione in tempo reale risulta necessaria per molte applicazioni di intelligenza artificiale, l'uso di immagini acquisite con frequenze elevate ma di risoluzione limitata diventa fondamentale per limitare il costo computazionale degli algoritmi. Anche in tal caso si evince una discrepanza tra fotogrammetria e CV, essendo nel primo caso impiegate immagini con risoluzione (sia geometrica che radiometrica) molto elevata, mentre nel secondo si è soliti usare direttamente video dalla risoluzione generalmente non superiore a 1,3 Mpixel. Considerando però il numero di immagini impiegate nei due casi rapportato alla risoluzione dei sensori, il numero complessivo di pixel utilizzati in entrambe le discipline potrebbe risultare piuttosto simile.
Un notevole vantaggio degli algoritmi di CV è che sono "non calibrati", ovvero non necessitano della calibrazione della camera (o videocamera) usata. Pertanto non è essenziale conoscere la focale, la posizione del punto principale e le dimensioni del sensore per ogni immagine. Da ciò risulta che lo zoom può essere impiegato senza alcun accorgimento, oppure possono essere impiegate contemporaneamente camere differenti senza nessuna conoscenza delle loro caratteristiche. La ricostruzione che si ottiene è però gravata da una ambiguità proiettiva, che comunque può essere rimossa mediante tecniche di *auto-calibrazione*.
Analizzando vantaggi e svantaggi reciproci di fotogrammetria e CV risulta spontaneo chiedersi se queste discipline possano integrarsi a vicenda e quali siano gli interessi comuni. Recenti lavori (Roncella, 2006) hanno dimostrato come sia possibile ovviare ad alcune problematiche dell'una mediante l'altra. Un esempio tipico è costituito dalla determinazione dei parametri approssimati per eseguire una triangolazione fotogrammetrica a stelle proiettive di un blocco terrestre, dove risulta vantaggiosa l'applicazione di procedure di CV.
L'algoritmo qui proposto permette di analizzare una sequenza di immagini a bassa risoluzione acquisite con una camera (o videocamera) digitale. L'algoritmo, in modo completamente automatico, determinerà una serie di punti omologhi per le coppie di immagini della sequenza (1-2, 2-3, 3-4, …) e, noti i parametri di calibrazione della camera, inizialmente calcolerà una ricostruzione metrica dell'oggetto per una coppia di immagini scelta come riferimento. Infine, tutte le restanti immagini verranno concatenate alla coppia di riferimento nella sequenza, aggiornando di volta in volta la ricostruzione. L'output è costituito dalle coordinate oggetto dei punti omologhi determinati in automatico e dai parametri di orientamento delle immagini.

1. La misura automatica dei punti omologhi tra le immagini della sequenza

L'algoritmo sviluppato può indistintamente analizzare singole immagini oppure un video. Nel secondo caso è necessario specificare la frequenza di campionamento del video per l'estrazione delle immagini da utilizzare, in quanto se queste fossero in numero eccessivo, il costo computazionale dell'algoritmo aumenterebbe moltissimo e contemporaneamente si avrebbe anche un peggioramento della geometria di presa. Tale frequenza deve dunque essere appositamente scelta dall'operatore considerando la frequenza di acquisizione della camera ma anche la velocità di spostamento della stessa, ricordando che l'algoritmo sviluppato ricerca gli omologhi tra un'immagine e quella successiva e che quindi si dovranno sempre ottenere coppie stereo.

La ricerca dei punti omologhi avviene per mezzo del *descrittore SIFT* (Lowe, 2004), che permette l'individuazione di features invarianti per scala e contemporaneamente esegue una classificazione delle stesse, in modo da facilitare le operazioni di individuazione degli omologhi sulle coppie.
Poiché solo utilizzando il descrittore SIFT è possibile determinare i punti omologhi tra coppie stereo di immagini spesso, seppur in numero limitato, vengono determinate false corrispondenze che devono essere opportunamente eliminate. Al tal fine, per ogni coppia l'algoritmo procede alla stima robusta delle geometria di presa mediante il calcolo della matrice fondamentale **F** (Hartley & Zisserman, 2001). Questa è una matrice 3×3 caratterizzata da rango 2 che descrive la geometria di presa di una coppia di immagini. Date alcune coppie di punti omologhi $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$ espresse in coordinate omogenee sulle due immagini della coppia stereo, la matrice fondamentale esprime il vincolo di geometria epipolare mediante la relazione:

$$\mathbf{x}_i'^T \mathbf{F} \mathbf{x}_i = \begin{pmatrix} x_i' & y_i' & 1 \end{pmatrix} \mathbf{F} \begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} = 0 \qquad (1)$$

La rimozione di false corrispondenze viene condotta mediante la stima robusta di **F** con la tecnica Least Median Squares (Rousseeuw & Leroy, 1987), che prevede la stima del minimo della mediana dei quadrati dei residui per l'insieme dei dati. Non esistendo una soluzione in forma chiusa, si fa ricorso a tecniche di *Random Resampling*, ovvero di estrazione di campioni casuali dall'intero set di dati costituiti dal numero minimo di dati necessario a calcolare i parametri. L'applicazione di questa tecnica robusta alla stima della matrice fondamentale avviene per mezzo della minimizzazione della distanza tra il punto sull'immagine e la retta epipolare calcolata a partire dall'omologo sull'altra immagine sempre per mezzo della matrice fondamentale (Scaioni, 2001). Infatti, le coppie di rette epipolari $\mathbf{l}_i' \leftrightarrow \mathbf{l}_i$ possono essere facilmente calcolate come $\mathbf{l}_i = \mathbf{F}^T \mathbf{x}_i'$ e $\mathbf{l}_i' = \mathbf{F}\mathbf{x}_i$, e quindi dipendono solo dalla matrice fondamentale determinata per ciascun campione estratto.
L'algoritmo sviluppato, implementato in parte in C ed in parte in MATLAB®, è in grado di analizzare una coppia di immagini in qualità VGA in meno di 5 secondi, fornendo le coordinate immagine degli omologhi per l'intera sequenza di immagini.

2. L'inizializzazione della struttura: le coordinate oggetto della prima coppia stereo

Scelta una coppia di immagini (tipicamente quella iniziale) è possibile ottenere le coordinate oggetto $\mathbf{X}_i$ dei punti omologhi $\mathbf{x}_i$ e $\mathbf{x}'_i$. La matrice fondamentale permette questa operazione mediante semplici algoritmi di triangolazione, ma fornisce una ricostruzione con una ambiguità proiettiva. La calibrazione della camera risolve completamente questo problema rendendo la ricostruzione sin da subito metrica (quindi con la sola ambiguità di scala) senza l'impiego di *auto-calibrazione*.
La calibrazione della camera può essere eseguita con qualsiasi software di calibrazione, ma in tal caso è stato utilizzato il *Camera Calibration Toolbox for Matlab* (Bouguet), che oltre ad essere gratuito mette a disposizione il codice di calcolo stesso.
La calibrazione fornisce la matrice di calibrazione **K** della camera (ed i parametri di distorsione), da cui possono essere determinate le *coordinate immagine normalizzate* per mezzo della sua inversa: $\hat{\mathbf{x}} = \mathbf{K}^{-1}\mathbf{x}$. Le coordinate normalizzate dei punti omologhi di una coppia stereo soddisfano l'equazione $\hat{\mathbf{x}}_i'^T \mathbf{E} \hat{\mathbf{x}}_i = 0$, del tutto analoga alla (1), dove **E** è detta *matrice essenziale* ed incapsula la geometria di presa di una coppia di immagini calibrate. I gradi di libertà di **E** sono dati dai 3 elementi di una matrice di rotazione e dalle 3 componenti del vettore che congiunge i centri di presa (simile all'orientamento relativo asimmetrico); con l'aggiunta di un ambiguità di scala questi parametri si riducono a 5.

Considerando che una camera è uno strumento che esegue una trasformazione del tipo **x=PX**, **x'=P'X** per una coppia stereo, dove **P** e **P′** sono matrici 3×4 dette "matrice camera", mediante un algoritmo di triangolazione e note che siano le camere è possibile stimare le coordinate oggetto $\mathbf{X}$.
Dalla matrice essenziale è possibile estrarre una coppia di camere per eseguire tale triangolazione. Generalmente si assume la prima camera del tipo $P = [I \mid \mathbf{0}]$ per determinare di conseguenza la seconda che avrà una forma del tipo $P' = [R \mid \mathbf{t}]$, ovvero data dalla matrice di rotazione della seconda camera e dal vettore congiungente i centri di presa. Tramite la decomposizione in valori singolari (SVD) di $E = UDV^T$ è possibile ottenere 4 soluzioni per la seconda camera P':

$$P' = [UWV^T \mid \mathbf{t}] \qquad P' = [UWV^T \mid -\mathbf{t}] \qquad P' = [UW^TV^T \mid \mathbf{t}] \qquad P' = [UW^TV^T \mid -\mathbf{t}] \tag{2}$$

dove, $t = U(0 \quad 0 \quad 1)^T$ e W è una matrice ortogonale del tipo:

$$W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

Delle 4 soluzioni solamente una è geometricamente corretta, ovvero rappresenta i punti davanti alle immagini. L'algoritmo quindi, ricevendo indicazioni su quale coppia di immagini debba essere utilizzata e noti i parametri di calibrazione della camera e le coordinate immagine, perviene ad una ricostruzione metrica per l'inizializzazione della sequenza.

3. Il concatenamento della sequenza

Le tecniche di geometria multi-immagine dimostrano che è possibile ottenere una ricostruzione 3D rigorosa usando coppie di immagini (mediante la matrice **F**/**E**), triplette (stima del *tensore trifocale*) oppure 4 immagini (tensore *quadri-focale*). Nel caso di 5 o più immagini (tipico di una sequenza) non esiste una soluzione con formulazione diretta come le precedenti, ma è necessario ricorrere ad un procedimento iterativo composto alternativamente da triangolazione e resezione.
In CV si definisce "Structure and Motion" (S&M) il processo di formazione di un modello 3D ("Structure") a partire dalle immagini dello stesso e dalle loro camere, opportunamente orientate ("Motion"). Nel presente lavoro, l'algoritmo aggiunge la terza immagine eseguendo una preliminare resezione per il calcolo della sua camera P'' mediante le coordinate oggetto $\mathbf{X}_j$ della prima coppia che compaiono anche sulla terza immagine ($\mathbf{x}''_j$). Infine aggiorna le coordinate oggetto aggiungendo quelle in comune alla coppia di immagini 2-3. La stessa procedura viene poi ripetuta per ogni immagine successiva (4, 5, …) sino al completamento della sequenza.

4. Esempi e sviluppi previsti

L'algoritmo è stato utilizzato per la ricostruzione di una facciata del Santuario di Re (VB), utilizzando un video di circa 10 secondi da cui sono state estratte 8 immagini acquisite con una Sony DSC-W30 con qualità VGA. Le 8 immagini sono risultate più che sufficienti per la ricostruzione dell'oggetto. In figura 1 sono riportate le prime due immagini della sequenza ed i punti omologhi estratti in modalità automatica mediante SIFT (in tutto 753) a cui sono state rimosse le false corrispondenze (dopo LMedS risultano 688 inliers). Successivamente sono state aggiunte le restanti immagini fornendo la ricostruzione proposta in figura 2, dove i punti segnalizzati sono stati aggiunti manualmente al solo fine di facilitare l'interpretazione e non sono stati utilizzati nei procedimenti di stima delle camere. Il processo di ricostruzione qui proposto è durato circa 40 secondi.

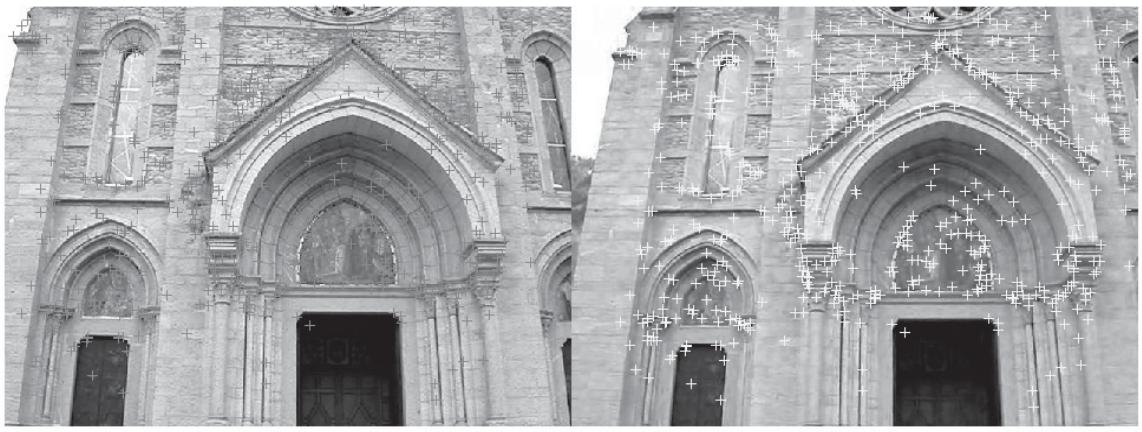

Figura 1 – Il matching automatico sulla prima coppia di immagini

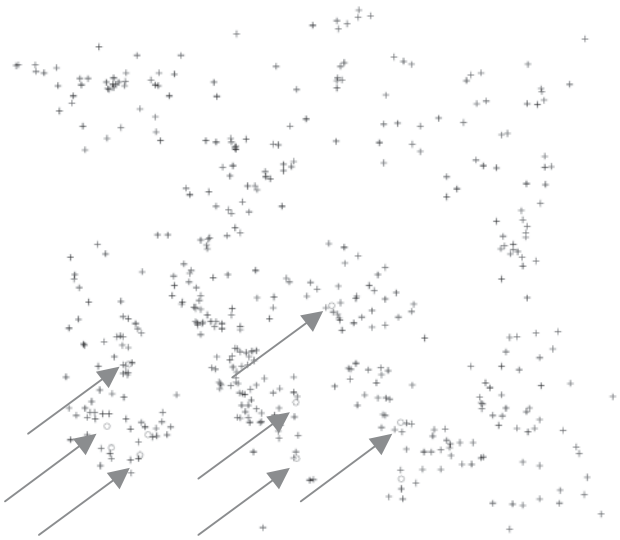

Figura 2 – La ricostruzione dalle 8 immagini. I punti segnalizzati sono stati aggiunti manualmente al solo fine di facilitare l'interpretazione

In un secondo esperimento è stato verificato il funzionamento dell'algoritmo quando la frequenza di campionamento del video è tale da generare un numero di immagini nettamente superiore a quelle strettamente necessarie, il che potrebbe causare un peggioramento della geometria di presa con angoli di intersezione di raggi per punti omologhi troppo acuti. Il video ha una durata di 10 secondi e sono state estratte 45 immagini. In tal caso l'algoritmo di matching automatico ha identificato un numero considerevole di corrispondenze corrette proprio per la minima differenza tra una immagine e la successiva (figura 3).

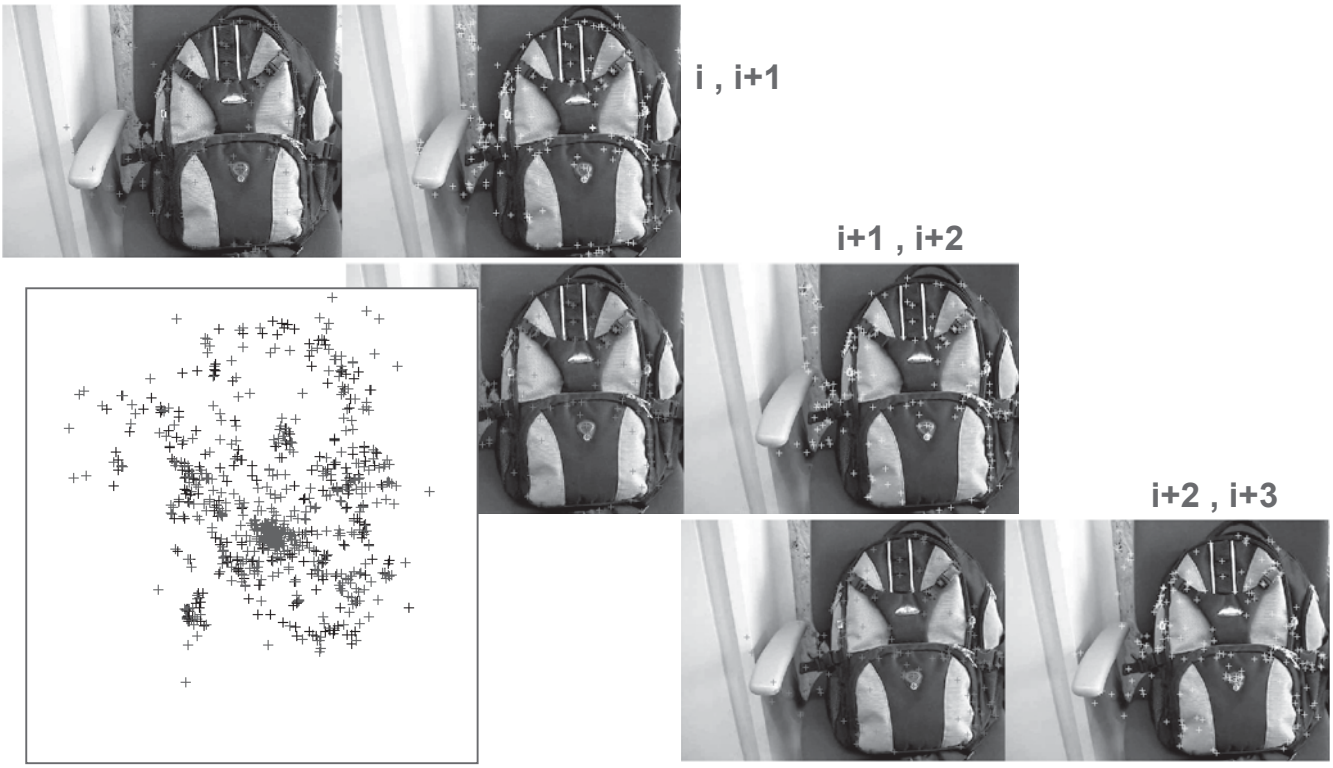

Figura 3 – Tre coppie di una sequenza di 45 immagini e la ricostruzione 3D finale

Il numero elevato di immagini ha permesso la determinazione di un maggior numero di coordinate oggetto, ma il modello creato non è ancora completo, come nel caso dell'esempio precedente. Per il completamento del lavoro sono possibili due scelte: la prima prevede l'aggiunta manuale di un numero di punti in zone significative misurati sulle immagini, in modo tale da completare la restituzione essendo noto l'orientamento delle immagini. La seconda invece è basata su algoritmi di *matching denso* per la determinazione automatica di una nuvola di punti su cui applicare la tessitura delle immagini per la formazione di un modello foto-realistico. Attualmente questa seconda soluzione è in fase di realizzazione.

Riferimenti bibliografici

Bouguet, J.Y. - *Camera Calibration Toolbox for Matlab*, www.vision.caltech.edu/bouguetj/calib_doc/

Hartley, R. I., Zisserman, A. W. (2001) - *Multiple View Geometry in Computer Vision*. Second Edition. Cambridge University Press, Cambridge (UK).

Lowe, D. G. (2004) - *Distinctive image features from scale-invariant keypoints*. *IJCV*, Vol. 60(2): 91-110.

Pollefeys, M., Koch, R., Vergauwenand, M., Van Gool, L. (2000) - *Automated reconstruction of 3D scenes from sequences of images*. *ISPRS JPRS,* Vol. 55(4): 251-267.

Roncella, R. (2006) - *Sviluppo e applicazioni di tecniche di automazione in fotogrammetria dei vicini*, Tesi di Dottorato in Ingegneria Civile, Università degli Studi di Parma.

Rousseeuw, P. J., Leroy, A. M. (1987) - *Robust regression & outlier detection*. John Wiley & Sons.

Scaioni, M. (2001) - *The Use of Least Median Squares for Outlier Rejection in Automatic Aerial Triangulation*. In: Proc. of the 1st Int. Symp. on "Robust Statistics and Fuzzy Techniques in Geodesy and GIS", Institute of Geodesy and Photogrammetry, Beright n. 295, Zurigo (Svizzera): 233-238.