Butlletí de la Societat Catalana de Matemàtiques
Vol. 31, núm. 2, 2016. Pàg. 179–193. DOI: 10.2436/20.2002.01.71

Fonaments geomètrics de la reconstrucció 3D

Joan Carles Naranjo

Resum: En aquest article fem una introducció als principis geomètrics que intervenen en la reconstrucció 3D d'una escena a partir de dues imatges digitals seves. Començarem explicant la modelització de la càmera i el seu calibratge. A continuació tractarem la geometria epipolar, és a dir, l'estudi de la posició relativa entre les dues càmeres des de les quals fem les fotografies de l'escena. Aquesta informació es codifica en la matriu fonamental, en el cas no calibrat, i en la matriu essencial, en el cas calibrat. Finalment descriurem com es realitza la reconstrucció efectiva a partir de la matriu essencial.

Paraules clau: càmera estenopeica, calibratge, matriu fonamental, matriu essencial.

Classificació MSC2010: 68T45.

1 Introducció

Des de fa temps ens estem acostumant que els ordinadors llegeixin per nosaltres. És ben habitual entrar en un pàrquing i que una càmera connectada a un ordinador i enfocada a la matrícula del cotxe «reconegui» les xifres i lletres que la componen. També sabem que moltes anàlisis clíniques estan automatitzades de manera que el comptatge de determinats tipus de cèl·lules no requereix la intervenció de l'ull humà. I no ens sorprenem quan ens diuen que s'ha pogut situar un robot petit sobre la superfície de Mart que es mou autònomament, decidint per ell mateix quina ruta ha de seguir i quins obstacles ha d'evitar; en definitiva, és un robot que «hi veu».

La llista d'exemples és interminable i cada vegada són més presents en la vida quotidiana. La branca de les ciències de la computació que se n'ocupa rep el nom de *visió per ordinador* (*computer vision*) i se sol considerar una part de l'àmbit de la intel·ligència artificial.

Del que volem parlar aquí és de les matemàtiques que s'utilitzen en alguns d'aquests processos. Com que el tema és amplíssim i només tenim la pretensió d'oferir-ne una pinzellada, farem dues reduccions dràstiques. La primera és que tan sols parlarem del que es coneix com a *reconstrucció 3D*, que, com fàcilment es pot imaginar, consisteix a extreure informació de dues o més fotografies digitals d'un mateix entorn per tal de poder deduir la posició en l'espai de tants punts com sigui possible de l'entorn original. La segona restricció, explicitada en el títol, consistirà a centrar-nos en els aspectes geomètrics del procés.

2 Una experiència docent

Abans de començar amb el contingut matemàtic voldria fer alguns comentaris. No sóc en absolut un expert en visió per ordinador, sinó que la meva relació amb aquesta àrea es redueix a participar en la docència d'una assignatura anomenada *geometria de la visió per ordinador* en la qual fèiem una introducció al tema posant l'èmfasi en la reconstrucció 3D. El que explicaré en les seccions posteriors és un resum en to divulgatiu de la meva part de l'assignatura: la geomètrica. Res no hauria tingut sentit si no hagués anat acompanyat de les altres dues parts: una introducció al processament d'imatges, a càrrec de José Ignacio Burgos, i unes pràctiques d'ordinador en llenguatge C ideades i implementades per Ferran Espuny amb la col·laboració del mateix José Ignacio Burgos. Tots dos varen aprofundir molt més que jo en aquesta àrea fins al punt que José Ignacio Burgos va acabar dirigint la tesi doctoral de Ferran Espuny en aquest tema. L'argument principal d'aquesta tesi [1], defensada el 2009, és el calibratge de les càmeres, un aspecte del qual parlarem de seguida.

El nostre interès per la visió per ordinador neix d'un seminari conjunt amb investigadors de l'IRI (Institut de Robòtica i Informàtica Industrial), dependent del CSIC, en el qual vam fixar com a objectiu explorar el llibre de Hartley-Zisserman [3]. Vam descobrir així un àmbit de recerca molt aplicat amb un component geomètric important, especialment projectiu, excel·lentment explicat en l'esmentat llibre i que ens va atreure immediatament. L'aplicabilitat i transversalitat del tema ens va fer creure que podria donar peu a una assignatura optativa (posteriorment també de màster) atractiva i útil per als estudiants de la Facultat de Matemàtiques de la Universitat de Barcelona, com així va ser. Diversos alumnes que la van cursar van continuar els seus estudis de visió per ordinador en altres centres del nostre entorn o de l'estranger i han acabat lligant la seva vida professional i acadèmica a aquesta àrea.

3 Una aproximació de butxaca a l'espai projectiu

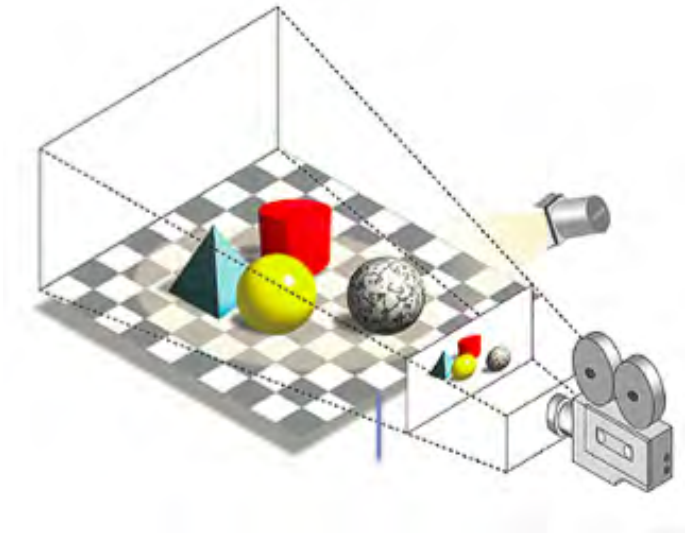
Com veurem en la propera secció, la modelització de les càmeres fotogràfiques se simplifica quan es considera en el context projectiu. Donem-ne una idea intuïtiva abans d'entrar en definicions: l'espai projectiu associat a $\mathbb{R}^n$ consisteix a afegir punts nous que corresponen a les «direccions de fuga» del nostre espai afí. En poques paraules, és un objecte geomètric on conviuen els punts i els vectors lliures com a elements del mateix conjunt. La manera més fàcil d'introduir-lo és utilitzant les anomenades *coordenades homogènies*. Posem-nos en el cas tridimensional ($n = 3$) per facilitar la notació. Un punt del nostre espai projectiu $\mathbb{P}^3$ està determinat per *quatre* constants reals ordenades que posem de la forma $(x : y : z : t)$ i que estan subjectes a dues normes: no totes elles són nul·les simultàniament, i dues quàdruples $(x : y : z : t)$ i $(x' : y' : z' : t')$ són considerades iguals (o que representen el mateix punt de $\mathbb{P}^3$) si existeix una constant de proporcionalitat λ de manera que $x' = \lambda x$, $y' = \lambda y$, $z' = \lambda z$ i $t' = \lambda t$. Així, si la darrera coordenada t és no nul·la podem dividir per ella i obtenir un element de la forma $(a : b : c : 1)$, en què les tres primeres coordenades estan normalitzades. Identificant (a, b, c) amb un punt de $\mathbb{R}^3$, tenim la igualtat de conjunts:

$$\mathbb{P}^3 = \mathbb{R}^3 \cup \{(x : y : z : t) \mid t = 0\}.$$

Aquesta darrera igualtat posa de manifest la idea expressada inicialment: l'espai projectiu $\mathbb{P}^3$ és una unió (disjunta) dels punts de l'espai afí ordinari $\mathbb{R}^3$ amb un pla d'equació $t = 0$ que parametritza, mòdul multiplicació per escalar, els vectors de l'espai. Així, la recta afí $(a, b, c) + \langle (u_1, u_2, u_3) \rangle$ s'identifica amb la recta de $\mathbb{P}^3$ que passa pels punts de coordenades homogènies $(a : b : c : 1)$ i $(u_1 : u_2 : u_3 : 0)$. En particular, dues rectes paral·leles de $\mathbb{R}^3$ passen a ser dues rectes secants i el punt d'intersecció es troba al pla de l'infinit.

4 Geometria d'una vista: càmeres i matrius

De manera molt simplificada es podria dir que una càmera fotogràfica consisteix en un aparell que en activar-lo els raigs de llum hi penetren dins per un forat minúscul i impacten en un pla. De manera ideal podem pensar, doncs, que una càmera és el parell format per un punt C per on passa la llum que es rep de l'escena, que anomenem *centre*, i per un pla $\mathcal{R}$ anomenat *retinal*. Aquest model s'anomena *model estenopeic* (*pinhole model*) o de *càmera obscura*.

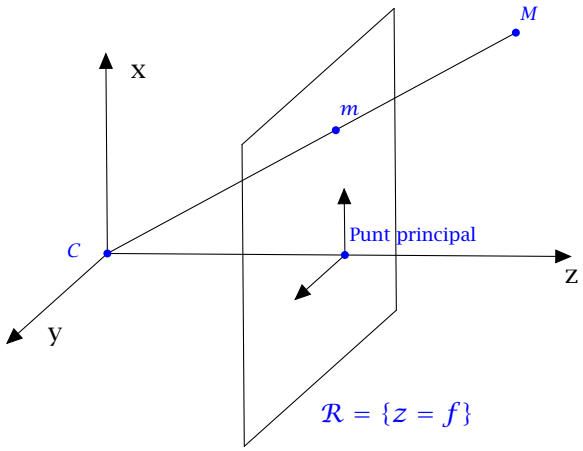


Suposarem que el pla està recobert per sensors digitals que tenen la capacitat de captar el color del raig de llum que hi ha impactat guardant-lo en la memòria de manera codificada, per exemple amb un codi RGB format per tres nombres, cadascun de 0 a 255, que ens indiquen la intensitat d'il·luminació vermella (R), verda (G) i blava (B) que cal posar per obtenir el color corresponent. Avui dia tots estem familiaritzats amb el terme *megapíxel* utilitzat en relació amb una càmera fotogràfica. Aquest terme ve a dir quants sensors tindrem en el pla retinal i, per tant, quina qualitat tindrà la nostra càmera (sense entrar en altres aspectes tant o més importants, com ara l'òptica): com més píxels més sensors i, per tant, més informació recollida per centímetre quadrat.

Per modelitzar geomètricament una càmera, la veiem com una aplicació

$$\mathbb{R}^3 \setminus \{C\} \longrightarrow \mathcal{R}$$

en què s'envia un punt M de l'espai (una mica pomposament se sol dir que és un punt *del món*) al punt del pla retinal m obtingut intersecant la recta MC amb $\mathcal{R}$.



Posem coordenades de manera que l'origen $(0,0,0)$ sigui el punt C i que el pla retinal $\mathcal{R}$ sigui el pla $z = f$, on f és la distància entre el centre i el pla retinal, que és el que s'anomena *distància focal*. Anomenem *punt principal* la projecció ortogonal del centre sobre el pla retinal; és el punt de coordenades $(0,0,f)$. Un càlcul senzill ens diu que si $M = (x,y,z)$, aleshores $m = (fx/z, fy/z, f)$. Trobem aquí una de les motivacions per passar-nos al càlcul amb coordenades homogènies o, dit d'una altra manera, per «projectivitzar» el problema: evitar quocients, convertir la transformació en lineal i, per tant, poder representar matricialment el nostre model. Considerem la càmera com l'aplicació:

$$\mathbb{P}^3 \setminus \{C\} \longrightarrow \overline{\mathcal{R}}$$

que envia $M = (x : y : z : t)$ a $m = (fx : fy : z)$. Tenim així una transformació governada per la matriu

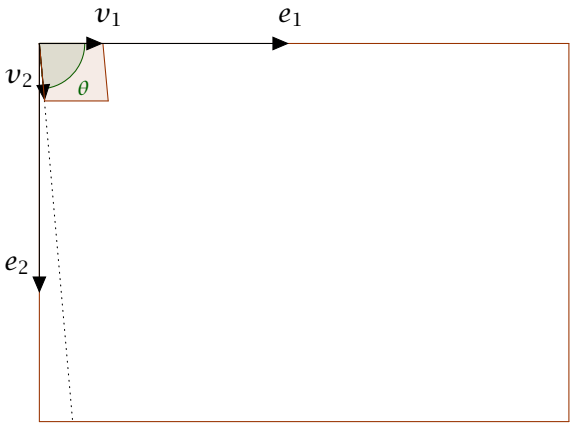
$$A = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

En realitat, les coses no són tan senzilles. Quan movem la càmera, les coordenades «món» van variant segons una rotació i una translació. D'altra banda, les coordenades escollides dins el pla retinal no són naturals; són molt simples des d'un punt de vista matemàtic, però poc ajustades a la manera com realment es poden portar a terme els càlculs. La informació que ens dóna una fotografia és una col·lecció de codis de colors cadascun del quals correspon a un píxel. Per tant, sembla més apropiat usar aquesta xarxa discreta de petitíssims quadradets que constitueixen la imatge. Per fer-ho agafem com a origen de coordenades l'extrem superior esquerre de la nostra graella de sensors, i com a eixos, les direccions que marquen els mateixos píxels. Les unitats les donen les longituds horitzontal i vertical del píxel. Aquest sistema de coordenades

no és necessàriament ortogonal, ja que, per molt perfecta que sigui la nostra càmera, l'angle θ entre els eixos pot ser lleugerament diferent de 90 graus, de la mateixa manera que els píxels poden no ser quadrats perfectes i, per tant, les unitats horitzontals i verticals no han de ser necessàriament iguals.

La conveniència d'usar aquest sistema de coordenades és fàcil d'explicar: suposem que hem fet dues fotografies amb una càmera digital d'una torre des de dues posicions diferents. Per fer una reconstrucció 3D en el nostre ordinador a partir de les dues imatges, el primer que necessitarem és reconèixer en cada imatge punts «que es corresponen», o sigui que provenen del mateix punt de la realitat. En particular, necessitarem referir-nos a punts concrets de cada imatge d'una manera manejable i quina altra manera millor hi ha que fer-ho dient, per exemple: el punt més alt de la torre ha sortit representat en la primera fotografia en el punt que està 3240 píxels cap a la dreta i 5329 cap avall?

Aquest canvi de coordenades comporta l'entrada en escena d'una de les matrius que tenen un paper més central en tot el problema: la *matriu de calibratge*. Consisteix en la matriu que transforma les coordenades que havíem pres inicialment en el pla $z = f$ en les coordenades píxel que hem considerat després.



En el dibuix superior, els vectors v_1 i v_2 són determinats per la forma del píxel, mentre que els vectors e_1 i e_2 corresponen a la base ortogonal que estem utilitzant a l'espai. Anomenem k_h i k_v el quocient entre la unitat de mesura que utilitzem (centímetres, mil·límetres…) i la grandària horitzontal, respectivament vertical, d'un píxel. Dit d'una altra manera, k_h i k_v ens diuen quants píxels caben en una unitat. L'angle θ mesura quant dista el píxel de ser rectangular (habitualment és gairebé de 90 graus). Obtenim fàcilment les fórmules:

$$\begin{aligned} e_1 &= k_h \cdot v_1, \\ \cos(\theta) \cdot e_1 + \sin(\theta) \cdot e_2 &= k_v \cdot v_2. \end{aligned}$$

Finalment, si (u_0, v_0) són les coordenades (píxel) del punt principal, trobem de les igualtats anteriors que la matriu del canvi de referència o matriu de calibratge és com segueix:

$$K = \begin{pmatrix} k_h & -\operatorname{cotan}(\theta) \cdot k_u & u_0 \\ 0 & \dfrac{k_v}{\operatorname{sin}(\theta)} & v_0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Dedicarem la propera secció a posar en relleu l'interès i les formes de càlcul de la matriu de calibratge. Ara, per cloure aquesta discussió sobre la modelització de la càmera estenopeica (*pinhole camera*), notem que la matriu associada a la càmera, amb la màxima generalitat possible, la podem escriure de la forma:

$$K \cdot A \cdot \overline{R},$$

on $\overline{R}$ és la matriu de quatre files i quatre columnes corresponent a una rotació R seguida d'un vector (columna) de translació t:

$$\overline{R} = \begin{pmatrix} R & | & t \\ 0 & | & 1 \end{pmatrix}.$$

5 Com podem saber com és per dins la càmera: un cop de martell?

Aturem-nos un moment en la informació que proporciona la matriu de calibratge que ens acaba d'aparèixer: els cinc paramètres k_h, k_v, θ, u_0, v_0 ens diuen com és la càmera per dintre. Fixem-nos que bàsicament ens indiquen la disposició i la forma dels sensors i, també, la posició relativa del centre i del pla retinal. Per això se'ls anomena *paràmetres interns*. Caldria afegir-hi la distància focal f, però la deixarem de banda atès que farem la nostra reconstrucció 3D sense atendre el factor d'escala. Esbrinar tots aquests valors és el que es coneix per *calibrar la càmera*. Per comoditat, suposarem que f és la nostra unitat de mesura, de manera que A és de la forma:

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

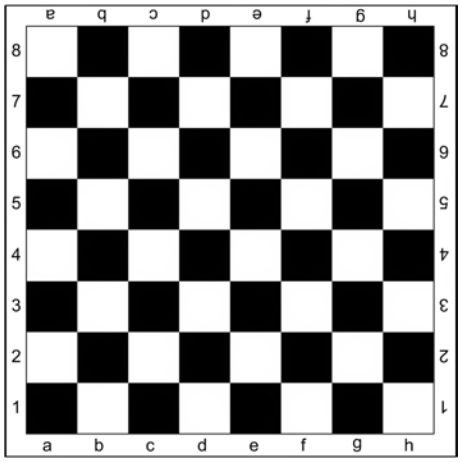
Ara podem simplificar per obtenir la matriu de tres files i quatre columnes

$$A \cdot \overline{R} = (R \mid t).$$

Així doncs, la matriu de la càmera es pot escriure de la forma:

$$K \cdot (R \mid t).$$

Però tornant al problema que hem plantejat: com podem calibrar la nostra càmera? Òbviament, el recurs de desmuntar-la de manera més o menys violenta (amb un tornavís o un cop de martell) és poc viable. La solució que s'aplica és força imaginativa. Tot i la seva complexitat quan es posa a la pràctica, se'n pot explicar la idea essencial en poques paraules. Comencem per triar objectes dels quals tinguem una informació mètrica molt precisa (en diem *patrons*) i en fem fotografies amb la nostra càmera. Els patrons possibles poden ser de molts tipus i s'ha de tenir present que seleccionar-ne uns o altres pot incidir en la precisió dels nostres resultats. Per fixar idees, pensem en un de ben simple i molt utilitzat: un tauler d'escacs amb caselles blanques i negres.



És una figura plana, la qual cosa introdueix un lligam en les coordenades espacials dels punts, conté molts parells de rectes formant angle recte i té molt contrast entre els dos costats de cada recta. Aquesta darrera propietat té un pes important en la part de la història que deliberadament estem amagant sota la catifa: el *processament de la imatge*, del qual parlarem una mica més avall. Suposem de moment que, un cop fetes les fotografies dels patrons, podem reconèixer en cada imatge quines coordenades píxel tenen els vèrtexs del tauler. A partir d'aquí, el mètode és relativament simple tot i que prou feixuc com per no descriure'l amb detall: diguem, simplement, que és fàcil saber quines propietats mètriques del tauler d'escacs físic són heretades per la seva imatge fotogràfica i que, en imposar-les, trobem les equacions que permeten calcular-ne els paràmetres interns i, per tant, calibrar la càmera.

Aquesta idea tan naïf està envoltada de dificultats que requereixen l'ús de tècniques molt variades. Per començar, si intentem obrir amb un editor una fotografia en format JPG, tindrem dificultats serioses, ja que és un format comprimit que s'utilitza per disminuir sensiblement la grandària del fitxer i que, per tant, perd informació. No podem aquí estendre'ns en aquest tema, però val la pena esmentar que aquesta compressió s'aconsegueix amb la intervenció decisiva de la transformada de Fourier en la seva versió discreta. Un cop descomprimida,

tindrem un fitxer de text ple de xifres agrupades de 9 en 9. Cadascun d'aquests blocs consta de tres enters que varien entre 0 i 255. Per exemple, si el fitxer comença per

$$000\,000\,000$$

voldrà dir que el píxel de la cantonada superior esquerra és negre, i si comença per

$$255\,255\,255,$$

el píxel serà blanc. La teoria del processament d'imatges ens ensenya a extreure informació a partir d'un fitxer com aquest. Una tècnica bàsica molt útil per al tipus de fotografia que tenim (no ho oblidem: un tauler d'escacs en blanc i negre) consisteix a restar a cada codi de color el codi del píxel precedent. D'aquesta manera, mirant els codis no nuls, detectarem aproximadament les línies que envolten les caselles. És un primer pas que dóna resultats molt minsos i que s'ha de completar amb moltes altres tècniques més elaborades per detectar formes, aïllar-les, segmentar-les... El lector curiós pot descobrir les bases d'aquesta vasta àrea de la computació donant un cop d'ull, per exemple, al llibre de Gonzalez-Woods [2].

Completada amb èxit l'anàlisi de la fotografia, disposarem de la informació aproximada que busquem. Salta a la vista que els errors són consubstancials al procés, en primer lloc, per la naturalesa discreta de les dades inicials. Així doncs, el càlcul dels nostres paràmetres intrínsecs estarà contaminat d'aquestes imprecisions. Per minimitzar l'error, realitzarem més fotografies del patró de manera que el nostre sistema d'equacions serà sobredeterminat. Finalment caldrà usar mètodes d'àlgebra lineal numèrica, majoritàriament basats en la descomposició SVD de les matrius, per trobar una solució òptima. No ens endinsem en cap d'aquests passos que hem descrit de manera tan superficial perquè el nostre objectiu és focalitzar-nos en la part geomètrica de la reconstrucció 3D, però esperem haver fet intuir la varietat de tècniques matemàtiques i de computació que intervenen en la solució d'aquest problema.

6 Dues vistes

Com veurem de seguida, la reconstrucció 3D es basa en dos pilars: conèixer les càmeres per dins i saber la posició relativa de les càmeres en el moment en què es fan les dues fotografies. Per tractar aquesta segona qüestió, ens trobem en una disjuntiva semblant a la del calibratge: usem la cinta mètrica, el transportador d'angles, un mesurador làser o qualsevol eina física més o menys moderna?; o, per contra, usem les fotografies fetes per la mateixa càmera per extreure la informació necessària per determinar la situació de les nostres càmeres? És clar que ens decantarem per la segona possibilitat, però és un camí que no està lliure d'algunes dificultats. Observem primer que el que pretenem calcular és de nou una informació matricial. En efecte, si posem com a origen de coordenades la primera càmera, o sigui, si suposem que la primera càmera té la matriu:

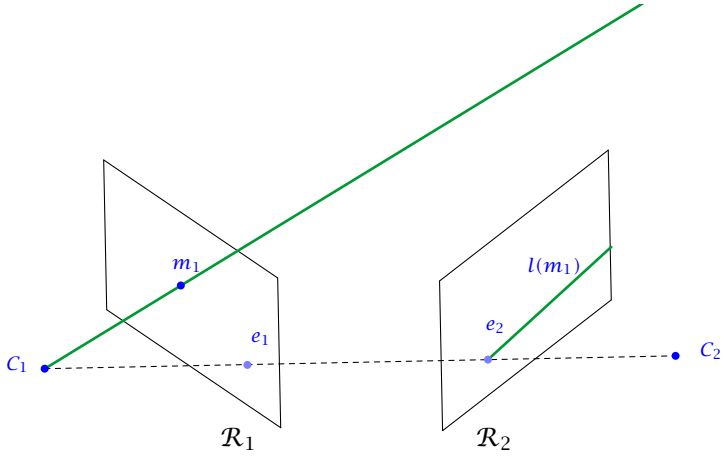
$$M_1 = K_1 \cdot (I_3 \mid 0)$$

(I_3 és la matriu identitat 3 per 3), aleshores la segona serà

$$M_2 = K_2 \cdot (R \mid t),$$

i la parella R, t és exactament la informació buscada: com cal rotar i traslladar la primera càmera per superposar-la a la segona.

El nostre objectiu en aquesta secció és descriure un pas intermedi en la recerca de R i t, que queda per a més endavant. Per fer-ho incorporem una nova definició que resulta molt intuïtiva: direm que dos punts m_1 i m_2, un de cada pla retinal, són *corresponents* si tots dos provenen del mateix punt del món. Per exemple, si fem dues fotografies d'un campanar i n'assenyalem a cada fotografia el punt més alt, estem seleccionant una parella de punts corresponents. En aquest exemple utilitzem la nostra capacitat per reconèixer objectes visualment i és just això el que volem evitar. Una mica de geometria ens pot ajudar a donar estructura als punts corresponents. Comencem per observar que, donat un punt m_1 en el primer pla retinal, hi ha molts punts potencialment corresponents amb ell en el segon pla. En efecte, unim m_1 amb el centre de la càmera i suposem, de manera ideal, que la recta així determinada es pot pintar d'algun color, com si fos un raig làser.



La fotografia amb la segona càmera de la recta acolorida dibuixa en el segon pla retinal una recta $l(m_1)$. Tenim així una aplicació que associa a les tres coordenades (projectives) del punt m_1 els tres coeficients de la recta $l(m_1)$ i que, per sort, és lineal. Així doncs, tenim una matriu de tres files i tres columnes F, anomenada *matriu fonamental*, que determina aquesta aplicació. Cal remarcar dues propietats importants per entendre i calcular F (e_1 i e_2 són, com s'indica en el dibuix, els talls de la recta C_1C_2 amb els dos plans):

a) La matriu té rang 2: com es pot observar en la figura superior totes les rectes $l(m_1)$ passen pel punt e_2 (anomenat *epipol*); per tant, no podem esperar exhaustivitat. De la mateixa manera, no és injectiva ja que tots els punts de la recta que passa per e_1 i m_1 tenen la mateixa recta $l(m_1)$ associada.

b) No costa gaire adonar-se que una parella de punts (m_1, m_2) són correspondents si, i només si, tenim la relació $m_2^T F m_1 = 0$.

Veurem que la matriu fonamental és la peça que permet trobar les matrius R i t, és a dir, la posició relativa entre les càmeres. És, doncs, important aprendre a calcular-la. Les propietats anteriors donen una idea de com es fa: suposem que tenim una col·lecció de punts corresponents proporcionada per alguna tècnica de processament d'imatges. Imposant la propietat b) tindrem un sistema d'equacions que permetrà calcular els coeficients de F. Un algorisme molt conegut és el dels *set punts*: s'imposa que set parelles de punts siguin corresponents. Com que la nostra matriu està determinada llevat de múltiple, arribem a una família de solucions possibles que depenen d'un paràmetre. Demanant que el determinant sigui zero obtenim un polinomi de tercer grau que, habitualment, només té una solució real i, per tant, determina F. Cal remarcar que un cop trobada la matriu fonamental tenim una eina addicional per al càlcul de parelles corresponents: d'una banda la propietat b) ens dóna una restricció que permet descartar correspondències falses, i de l'altra, es pot usar F per *rectificar* les fotografies de manera que les línies epipolars $l(m_1)$ siguin totes horitzontals, la qual cosa simplifica la localització de parelles de punts corresponents.

Finalment, abans de cloure la secció, un esment de la dificultat de trobar punts corresponents. Hi ha molts mètodes i, sovint, les tècniques més eficients són combinacions de molts d'ells. Una idea molt simple però que dóna una primera aproximació al problema és la següent: considerem finestres petites de píxels dins de cada fotografia; per exemple, quadrats de 5 per 5 píxels. Cada quadrat es pot veure com una matriu en la qual les entrades són els codis RGB dels píxels. Prenent alguna noció de distància entre matrius, podem provar d'identificar quadrats molt propers d'una i altra fotografia. Si la distància és realment molt petita, això ens indica que possiblement els centres dels quadrats es corresponen.

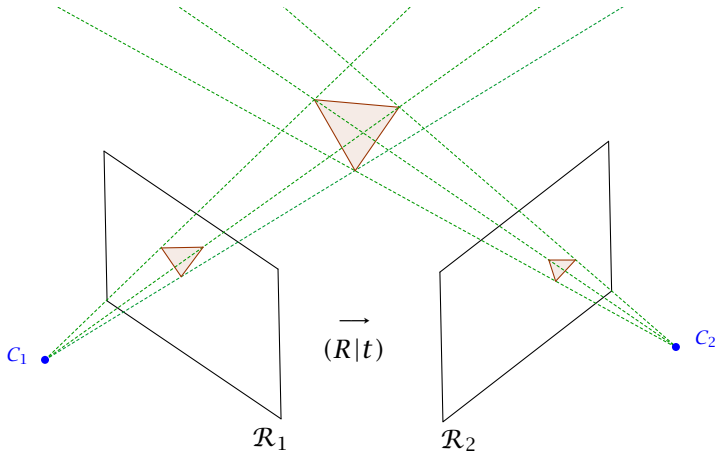
7 Reconstrucció 3D

En aquesta secció explicarem com es porta a terme la reconstrucció partint de les eines introduïdes en les darreres seccions: suposem que hem sabut calibrar les càmeres i, per tant, que coneixem les matrius K_1 i K_2, que hem calculat també de la matriu fonamental F i, finalment, que disposem d'una bona col·lecció de parelles de punts corresponents $P = \{(m_{1,i}, m_{2,i})\}$. Abans de dir què podem fer amb tota aquesta informació, un comentari sobre el conjunt P: la facilitat per obtenir una col·lecció P amb molts punts i de molta «qualitat» dependrà del tipus d'entorn que intentem reconstruir. Òbviament, dues fotografies del mateix cel blau ens donaran poques oportunitats per trobar parelles significatives, i la reconstrucció no tindrà cap sentit. De la mateixa manera, pot passar que un punt del món surti en una de les fotografies i no en l'altra i, per tant, no originarà una parella de punts corresponents. Salta a la vista que el nostre objectiu serà més complicat en fotografies fetes a la natura que les que es fan en l'entorn urbà o en un espai tancat on abunden les línies rectes, els canvis molt contrastats de colors i els volums més geomètrics (de vegades es diu que és un *entorn estructurat*).

Tornant al nostre discurs inicial, ja hem comentat que podem suposar que les matrius que modelitzen les nostres càmeres són de la forma:

$$M_1 = K_1 \cdot (I_3 \mid 0) \quad \text{i} \quad M_2 = K_2 \cdot (R \mid t).$$

Si realment coneixem les dues càmeres, o sigui, si obtenim K_1, K_2, R i t, podem esperar que els punts del món P dels quals provenen les nostres parelles es puguin calcular. En efecte, conèixer les matrius de calibratge equival a saber on està posat el centre de cada càmera en relació amb els plans retinals corresponents. A més a més, R i t ens donen la posició relativa entre les dues càmeres; per tant, podem passar un raig de llum pel centre de la primera càmera i un punt $m_{1,i}$ i un altre raig de llum pel centre de la segona càmera i per $m_{2,i}$, i obtenim dues rectes que es tallaran en el punt que estem buscant. Bé, en realitat el més probable, a causa dels errors acumulats, és que no es tallin sinó que passin molt a prop, de manera que caldria trobar els punts de cada raig que minimitzen la distància i prendre el punt mitjà.



En realitat el que acabem de dir pretén donar la idea de per què la informació K_1, K_2, R, t és la necessària per a la reconstrucció 3D; però, pel que fa al càlcul, el plantejament és lleugerament diferent: el punt X_i del món que correspon al parell $(m_{1,i}, m_{2,i})$ és la solució de

$$M_1 \cdot X_i = m_{1,i},$$
$$M_2 \cdot X_i = m_{2,i}$$

(les coordenades de X_i s'han d'escriure «normalitzades», o sigui, $X_i = (x_i, y_i, z_i, 1)$). Utilitzant tècniques d'àlgebra lineal numèrica obtenim, sempre aproximadament, les coordenades del punt buscat.

Com a resum d'aquesta secció, podem dir que el problema de la reconstrucció queda reduït al càlcul de les matrius R i t a partir de les dades K_1, K_2 i F. Aquest és el contingut de la propera secció.

8 La matriu essencial i la seva descomposició

En arribar a aquest punt hem convertit el problema de la reconstrucció 3D en un problema purament matricial: gràcies a una combinació de mètodes molt variats hem aconseguit disposar de les matrius K_1, K_2, F (aquesta darrera determinada llevat de multiplicar per una constant) i una col·lecció de parelles de punts corresponents. Segons el que hem explicat en la darrera secció, per acabar només ens queda calcular la matriu de rotació R i el vector de translació t. El resultat crucial en aquesta part de la teoria és la relació següent (entre matrius de tres files i tres columnes de rang 2):

$$K_2^T F K_1 = [t]_x \cdot R,$$

on $[t]_x$ és una matriu antisimètrica construïda a partir del vector de translació $t = (t_1, t_2, t_3)$ de la forma següent:

$$[t]_x = \begin{pmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_1 \\ -t_2 & t_1 & 0 \end{pmatrix}.$$

Aquesta relació posa de manifest de manera explícita el que ja havíem avançat: la matriu fonamental codifica la posició relativa entre les càmeres. Val a dir que la matriu $E := K_2^T F K_1$ que apareix a la relació rep a la literatura el nom de *matriu essencial*. Observem que és calculable a partir de les nostres dades i que el que volem és descompondre E en el producte $[t]_x \cdot R$. Justifiquem ara que això es pot fer de manera única, llevat de signe. O sigui, en realitat hi ha quatre descomposicions possibles de les quals només una és la que ens convé. No és gaire difícil eliminar *a posteriori* aquesta indeterminació: si agafem un signe inapropiat per a R o per a t, ho notem en el fet que el conjunt de punts P que usem per a la reconstrucció 3D queda situat «al mateix costat» que el centre d'alguna de les càmeres en relació amb el pla retinal.

Recuperació de *t*

Atès que la matriu E està determinada llevat de la multiplicació per una constant, comencem per normalitzar-la d'alguna manera. Imposem que la norma de la matriu sigui 1, $\|E\| = 1$. Usem com a norma l'euclidiana de $\mathbb{R}^9$ després d'identificar la matriu amb un punt d'aquest espai. És a dir, estem imposant que la suma dels quadrats de les entrades de la matriu sigui 1. Així E quedarà fixada llevat de signe. Aquesta norma de matrius que estem considerant té propietats força agradables. Per exemple, és fàcil comprovar que la norma al quadrat d'una matriu A és la traça (suma dels elements de la diagonal) de la matriu $A \cdot A^T$. En particular, si suposem E ja descomposta, $E = [t]_x \cdot R$, tenim:

$$\begin{aligned} \|E\|^2 &= \text{traça}(E \cdot E^T) = \text{traça}([t]_x \cdot R \cdot R^T \cdot [t]_x^T) = \\ &= \text{traça}([t]_x \cdot [t]_x^T) = \|[t]_x\|^2 = 2\|t\|^2. \end{aligned}$$

Per tant, fixar la norma de la matriu essencial equival a fixar la norma del vector de translació $\|t\|^2 = \frac{1}{2}$. Per un altre costat observem que:

$$E^T = R^T \cdot [t]_x^T = -R^T \cdot [t]_x.$$

Atès que la matriu $[t]_x$ té nucli $\langle t \rangle$ i R té determinant no nul, el nucli de E^T i la condició sobre la norma ens proporcionen, llevat de múltiple, el vector de translació t.

Unicitat llevat de signe de la descomposició

Un cop el vector de translació t és conegut, la matriu R queda determinada per les equacions:

$$E = [t]_x \cdot R,$$
$$R \cdot R^T = I_3.$$

Notem que no podem «aïllar» R en funció de E i $[t]_x$ perquè la segona matriu no és invertible. Aquest sistema té una única solució (deixant de banda el signe). La demostració es basa en un bonic argument geomètric que volem compartir amb el lector que ha aconseguit arribar fins aquí. Suposem que tenim:

$$E = [t]_x \cdot R = [t]_x \cdot R',$$

on R' és una altra rotació. Aleshores, multiplicant per la dreta amb la matriu transposada de R' tenim

$$[t]_x \cdot R \cdot R'^T = [t]_x.$$

Transposant tota la igualtat, ens queda que la rotació $R_0 = R' \cdot R^T$ satisfà:

$$R_0 \cdot [t]_x = [t]_x.$$

Hi ha una característica de les matrius $[t]_x$ en la qual no hem incidit fins ara i que és molt fàcil de comprovar usant la seva definició: per a qualsevol vector de l'espai u, la imatge $[t]_x(u)$ és simplement el producte vectorial $t \wedge u$. Per tant, la imatge de l'aplicació amb matriu $[t]_x$ és el subespai vectorial de dimensió 2 format pels vectors ortogonals a t. La igualtat obtinguda més amunt ens diu que R_0 deixa invariants (són vectors propis de valor propi 1) tots els vectors ortogonals a t. Atès que una rotació de l'espai que no sigui la identitat només deixa invariants els vectors d'un subespai de dimensió 1 (el seu eix), tenim que $R_0 = I_3$ i, per tant, $R = R'$.

9 Resum i conclusió

Si fem una llista ordenada dels processos que han intervingut en el mètode explicat per a la reconstrucció 3D, trobem:

a) En primer lloc, calibrem les càmeres usant fotografies de patrons.
b) A continuació, localitzem amb tècniques de processament d'imatges algunes parelles de punts corresponents.
c) Aquestes parelles de punts ens permeten calcular la matriu fonamental mitjançant, per exemple, el mètode dels set punts.
d) Usant la informació proporcionada per la matriu fonamental, rectifiquem les fotografies i calculem més parelles de punts corresponents.
e) Calculem la matriu essencial i la seva descomposició en rotació i translació.
f) Calculem els punts del món que corresponen a les parelles de punts corresponents.

Aquesta llista de procediments que barregen geometria lineal i projectiva, àlgebra lineal numèrica, mètodes d'optimització, processament d'imatges i programació, entre d'altres, dóna una idea de la varietat i transversalitat a la qual fèiem referència a l'inici d'aquestes notes.

NOTA. Les figures de les pàgines 183, 188 i 190 han estat realitzades amb el programa Geogebra.

Agraïments

Vull agrair afectuosament a Anna Puig i a Ferran Espuny la lectura d'una versió prèvia d'aquestes notes. Les seves aportacions han millorat notablement el resultat final.

Referències

[1] ESPUNY, F. *Self-Calibration of Projective and Generic Central Cameras*. Tesi doctoral. Universitat de Barcelona, 2009.

[2] GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing*. 3a ed. Upper Saddle River, NJ: Prentice Hall, 2008.

[3] HARTLEY, R.; ZISSERMAN, A. *Multiple View Geometry in Computer Vision*. 2a ed. Cambridge: Cambridge University Press, 2004.

DEPARTAMENT DE MATEMÀTIQUES I INFORMÀTICA
FACULTAT DE MATEMÀTIQUES I INFORMÀTICA
UNIVERSITAT DE BARCELONA
jcnaranjo@ub.edu