Válaszok Dr. Hajdu András bírálatára

1. Általános köszönetnyilvánítás

Válaszaim elején mindhárom bírálónak közösen szeretném megköszönni az alapos munkájukat. Miután a számítógépes látás területe hatalmas, az én saját témáimmal rajtam ás tanítványaimon kívül nem sokan foglalkoznak, pláne nem MTA doktora címmel rendelkező tudósok. Ennek ellenére a bírálatok nemcsak arról tanúskodnak, hogy megértették a meglehetősen tömény leírást, hanem mindhárman adtak olyan ötleteket, amelyek további vizsgálatokra sarkalnak, érdekes kérdéseket vetnek fel, amelyre érdemes lenne megtalálni a választ.

A tisztelt Bírálók motiváló megjegyzéseinek kézzelfogható eredménye is volt, mivel az Arxiv preprint szerverre két munkát is feltettem [Haj26a, Haj26b], melyeket később hivatalos fórumokra is szeretnék beküldeni.

2. Kérdések a 2. fejezethez

A geometriai és a tanulás-alapú módszerek együttélésében melyek azok a problémák, ahol a klasszikus geometriai formalizmus ma is nélkülözhetetlen előnyt ad (pontosság, robusztusság, interpretálhatóság)?

Erre a kérdésre két okból is nehéz válaszolni.

1. Egyrészt azért, mert nehezebb éles határt húzni a geometriai módszerekkel, illetve gépi tanulással hatékonyan megoldható feladatok között, mint azt első látásra gondolnánk.

2. Másrészről pedig azért, mert sajnos a publikált módszerek jelentős részét nem lehet reprodukálni, azaz nem megbízhatóak. Azaz a magam részéről minden gépi tanulással elért eredményt megkettőzött gyanakvással figyelek. Ami nem jelent azt, hogy nem ismerném el az áttörő jelentőségét a gépi tanulásnak.

   A számítógépes látásban szerintem a reprodukálhatóság [1] legnagyobb ellensége nem a szándékos csalás, hanem a folyamatos sürgetés: a kutatóknak muszáj jobb eredményt felmutatniuk, mint az előző cikk, ezért (tudatosan vagy tudat alatt) úgy állítják be a kísérletet, hogy az az ő gépükön és az ő adataikon fejlődést mutasson, viszont arra nincsen energiájuk, hogy sokféle konfiguráción teszteljék a módszereket. Konfiguráció alatt értem azt is, hogy más gépen is kellene futtatni, de azt is, hogy másfajta adatbázison nem próbálják ki, és az ő módszerüket rátanítják a kiválasztott adatokra.

De nem szeretnék a kérdés elől kitérni. Ha egy feladat geometriailag nem leírható, mint például egy állat alakja: az egyedek különbözősége, a képkészítés helyének kiszámíthatatlansága lehetetlenné teszi az állat pontos modellezését. Vagy egy önvezető jármű esetén az úton átfutó "gyalogos" lehet egy kerékpárt toló embertől kezdve a bottal közlekedő öreg nénin át az ikerbabakocsiig sok minden. De még a sávdetektálás is ilyen, mert sz országok közötti különbözőség, a festék kopása, az úthibák mind-mind egyediséget képesek okozni. Ezeket szinte lehetetlen geometriai modellel megadni, ezért ilyenkor véleményem szerint a gépi tanulás megkerülhetetlen.

[1] "Szabó Csaba: Elpazarolt orvostudomány - hiteltelen kutatók, megbízhatatlan kísérletek" című munkája jól bemutatja a jelenséget a kiváltó okokkal együtt egy tőlünk szerencsére távol lévő tudományterületen. Személy szerint azért attól félek, hogy a kísértések nemcsak a biokémiával foglalkozó kutatók között léteznek, sőt, a mesterséges intelligencia megjelenésével felerősödnek az informatikai területeken is.

A tisztán geometriai problémák közül a dolgozatomban is említett feladatok közül a 3D rekonstrukciót (stucture from motion) vagy az offline kalibrációt említeném. A rekonstrukcióban viszont csak a kamerák külső kalibrációját (pózbecslés: forgatás és eltolás számítása) érdemes venni, a 3D pontfelhő nagyon ritka lesz, azt a mélytanulás módszereivel (például monodepth1[YKH+24] vagy a legújabb DUSt3R háló [WLC+24]) sokkal sűrűbbé lehet tenni.

Azonban létezhet a két megközelítés egyesítése: amikor a geometriai törvényszerűségeket be lehet építeni egy tanulás során a hibafüggvénybe ('loss function') Itt meg is szeretném jegyezni, hogy a saját eredményként közzétett III. törvényt (I.2. tézispont), amely az alapvető (fundamentális) mátrix és az affin transzformáció kapcsolatára ad egy 3D egyenletet, kollégák már sikeresen beillesztették tanuló algoritmusba [SGY+25]. Mi magunk is foglalkozunk azzal – egyelőre publikálatlan eredményről van szó–, hogy a normálvektorokra felírt első törvényt bevonjuk egy tanulási folyamat költségfüggvényébe.

A későbbi fejezetekben központi szerepet kap az affin transzformáció megbízható becslése. Mely tényezőt tartja kritikusabbnak: az affin információt előállító lokális detektorok pontosságát, vagy a geometriai solver-ek numerikus stabilitását?

Miután alapvetően 3D látással foglalkozom, és az affin transzformációk becslése kétdimenziós feladat, egyértelműen abban vagyok bizonytalanabb, hogy hogyan és milyen minőségben lehet kinyerni a képekből a szükséges információt. A megoldók pontosságára könnyen lehet kvantitatív kiértékelést készíteni, az affin transzformációk pontosságának ellenőrzésére nehezebb ilyen vizsgálatot elvégezni.

Ennek ellenére már tettem én is kísérletet, hogy megbecsüljem, pontosabban fogalmazva számszerűsítsem az affin transzformációk pontosságát. Furcsa leírni, hogy ezeket az eredményeket nem publikáltam, a Computer Vision and Image Understanding folyóirat speciális számába elküldtem, de nem fogadták el, és aztán elfeledkeztem róla. Egymásra merőleges sakktáblákat használtam, hogy alapigazság (ground truth) értékeket állítsak elő annak eldöntésére, hogy milyen pontos becslő eljárásokat lehet alapozni affin transzformációkra. Most ennek a válasznak a kiegészítéseként feltettem az Arxiv preprint szerverre [Haj26a]. Ide nem másolnám be a teljes munkát terjedelmi okok miatt.

Hogy ne hagyjam érdemi, de rövidebb válasz nélkül a Bírálót, egy egyszerűsített vizsgálatot is végeztem. Ennek alapja, hogy amennyiben a kamera mozgása szabályos, az affin transzformációk egyes paramétereit ismerhetjük. Például, ha egy folyamatosan forgó drónra rögzített kamera a talajt nézi, a kamera tengelye pontosan függőleges, akkor két képkocka között az affin transzformációknak egyszerű 2D forgást kell leírniuk. Az ettől való eltérést pedig már lehet számszerűsíteni.

Mostanában, ahogyan azt Szirmay-Kalos Lászlónak adott válaszomban is leírtam (ennél kicsit részletesebben), a SIFT jellegzetes pontokból [Low04] kiindulva az "ősrégi" Lucas-Kanade iteratív algoritmust [LK81, BM04] használom a mintaillesztés pontosítására, az alkalmazott modell pedig a hatparaméteres affin transzformáció. Érdekességképpen ORB [RRKB11] jellegzetes pontokra is megvizsgáltam az eredményt.

A módszer a következő volt: detektáltam jellegzetes pontokat a két képen, és az N legjobb megfeleltetést kiemeltem [2]. Mind a SIFT, mind az ORB leírókhoz forgatás (orientáció) és skálázás tartozik. Ebből egy kezdeti affin transzformáció számítható, azzal a megkötéssel, hogy ez csak két paraméteres lesz, hiszen nyírás egyáltalán nincsen, és a vízszintes és függőleges skálázás meg fog egyezni.

Az így kapott transzformációkat a Lucas-Kanade iteratív algoritmusnak beadtam kezdeti értékként, és ez már az összes paramétert, az eltolást is beleértve, kiszámolta.

Az eredmények az 1. táblázatban láthatóak. Két olyan képkockát választottam, ahol kicsi a forgatás (körülbelül kettő fok), hogy a nagyon rossz értékek szembetűnőbbek legyenek. Az algoritmusoktól 50, 100, 250 és 500 pontot kértem, az ORB azonban ötszázat már nem tudott kinyerni, ezért szerepel kevesebb adat ennél a módszernél.

A 250 pontot kérő módszerek eredményeit az 1. ábrán is lehet látni. Ez tovább erősíti a képet, hogy a SIFT leírók alkalmasabbak a feladatra, hiszen sokkal egyenletesebb az eloszlása a pontoknak, az ORB esetén a pontok zöme a jobb felső sarokban található, a SIFT esetén a kép közepén is található minta.

Ezek után a kapott affin transzformációkat dolgoztam fel. A drón speciális mozgásából adódik, hogy egy 2D-s forgatást kellene leírnia a 2×2-es transzformációnak. Ezért a kapott affin mátrixhoz megkerestem legkisebb négyzetes értelemben a legközelebbi 2D forgatást, és annak a szögét vettem forgatási értéknek. Mind az N ponthoz kaptam ezért értéket.

[2] A detekcióhoz és megfeleltetéshez az OpenCV algoritmusait használtam.

		Medián alkalmazása		Kimerítő keresés	
Módszer	Detektál pontok (db.)	Inlier (%)	Szög (◦)	Inlier (%)	Szög (◦)
SIFT	50	56	1,8284	58	1.8609
	100	46	1,8120	50	1,8749
	250	39,2	1,6849	41,6	1,5791
	500	31,4	1,7275	31,8	2,0309
SIFT+LK	50	90	1,8851	92	1,8517
	100	88	1,8901	89	1,8838
	250	89.2	1,8839	90	1,8744
	500	91.6	1,8819	92	1,8794
ORB	50	28	1,3128	38	1,8385
	100	26	1,4638	29	1,8005
	250	22.4	2,0127	24.4	1,8068
ORB+LK	50	92	1,8033	92	1,8113
	100	96	1,8186	96	1,8284
	250	93.6	1,8456	93.6	1,8333

1. táblázat. Mérési eredmények szabályosan forgó drón felvételére. A cél a forgás robusztus becslése. A táblázatok a végső inlierek arányát és a becsült forgatási szöget tartalmazzák különböző módszerekre, különböző mintaszámmal.

Kétféle módszerrel végeztem el a forgatás robusztus becslését. Először ezeknek az értékeknek a mediánját vettem végső forgatásnak. A belső értékeket (inlier) pedig úgy számoltam meg, hogy egy kis küszöbértékkel a medián alatt vagy felett kellett lennie a kapott szögeknek. A küszöböt $0,5°$-nak választottam. A második robusztus módszer egy kimerítő keresést végzett: az összes kapott szöghöz képest, ugyanazzal a fél fokos küszöbbel meghatározza a jó pontokat (inlierek), és ezeknek az átlaga lesz az eredmény. Ez kiküszöböli a mediánnak azt a hátrányát, hogy csak akkor működik, ha a pontok fele jó érték.

A táblázat értékeléséből, a teljesség igénye nélkül, az alábbi következtetéseket lehet levonni:

- A SIFT detektor alkalmazása esetében a kezdeti becslés pontosabb, mint ORB alkalmazásakor.
- a kimerítő keresés értelemszerűen hatékonyabb a medián alkalmazásánál, cserébe lényegesen lassabb is, bár futási időt itt most nem vizsgáltam.
- A Lucas-Kanade algoritmus [LK81] rendkívül hasznos, mind az inlier arányt sikerült szignifikánsan megemelnie, és a kapott értékek is az $1,8 \ldots 1,9$ fokos intervallumba esnek.

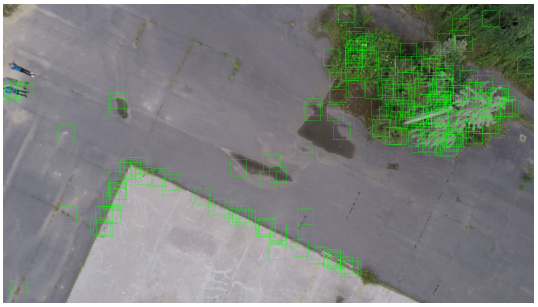
3. Kérdések a 3. fejezethez

A szerző a lokális affinitást differenciális (elsőrendű) geometriai információként használja. Mennyiben tekinthető ez a felületi lokális struktúra közvetett „mérésének", és a gyakorlatban mennyire stabil ez az információ zajos képi környezetben?

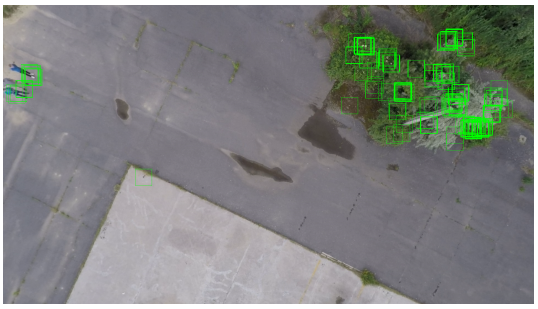
Az előzőekben már említett Arxiv preprint [Haj26a] tartalmazza, hogy sakktáblák esetén milyen pontossággal lehet meghatározni az affin transzformációt. A sakktábláknak a sarkait vettem kiindulási pontoknak, és az élirányokból lehet az affin transzformációt megbecsülni a II.9. tézispontban javasolt módokon. Ebből a preprintből emelnék ki egy eredményt.

Az összehasonlított módszerek az alábbiak:

- **F2UDIR**. Az affin transzformációt a ismert fundamentális mátrix alapján becslültem meg, és a vízszintes és a függőleges irányokat felhasználva.
- **F3UDIR**. Hasonló az F2DIR-hez, azonban a becslés során három irányt veszek figyelembe, beleértve az átlós irányt is. A probléma ekkor túlhatározott, mivel két nem skálázott irány elegendő az affín becsléshez, ha az epipoláris geometria ismert.




(a) 250 darab SIFT [Low04] jellegzetes pont követése




(b) 250 darab ORB [RRKB11] jellegzetes pont követése

1. ábra. Forgó drónról készült felvételek és a detektált minták az első képen (bal), és a becsült affin transzformáció (jobb)

- **DET3UDIR**. A becsléshez csak három nem skálázott irányt használok. A skálát az affin transzformációk determinánsa határozza meg, miután tudjuk, hogy a minta területének változása az affin transzformáció determinánsával egyezik meg [3].
- **2SDIR**. Két skálázott irányt, a vízszintest és a függőlegest veszem figyelembe.
- **3SDIR**. Három skálázott irányt – beleértve az átlóst irányt is – veszem figyelembe, így a probléma túlhatározott.

Az eredményeket a 2. ábrán lehet megtekinteni. Az élirányokra a teljes kiértékelést itt most mellőzném, de az jól látszik, hogy skálázás figyelembe vételével lehet 10 fok alá szorítani a hibát, ha 3 fokos a (sakktábla mezőinek) kétdimenziós irányokat érintő nulla várható értékű normális zaj szórása. Ez az érték már egészen komoly hibának minősül, ennél kisebb értékeket is el lehet érni, lásd a fenti, drónos felvételekből származó elforgatási hibát.

További vizsgálati eredmények, beleértve valós sakktáblás teszteket, a preprint anyagban [Haj26a] olvashatóak.

A levezetett összefüggések mennyiben általánosíthatók nem perspektív kameramodellekre (pl. halszem/omnidirekcionális), illetve milyen fő nehézségek jelentkeznek ebben?

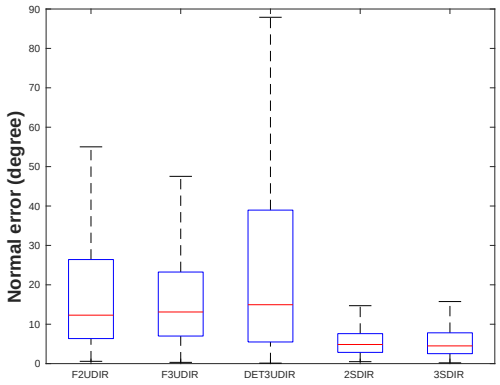
Az I. törvény általános kameramodellek esetén is igaz. Viszont az összes többi levezetés kizárólag lyukkamerára (pin-hole camera, camera obscura) lett levezetve. Ez nem jelent hatalmas megkötést, mert a legtöbb kamera középpontos vetítést ad, így a képeket "ki lehet egyenesíteni", és a lyukkamera modelljét lehet a továbbiakban alkalmazni [4].

Megpróbálkoztam még egy általános affin kamerával [5] is, amikor a homogén osztást kiküszöböljük,

[3] Ez a megállapítás már túlmutat a II.9. tézisen, azóta fejlesztettem ki

[4] Például az OpenCV könyvtárban az *undistort()* függvény végzi el ezt a feladatot.

[5] Az affin kamerának nincsen köze a disszertációban taglalt affin transzformációkhoz, szerencsétlen névegyezéssel van dolgunk. Affin kamera esetén $\mathbf{p}_{2D} = [u \quad v]^T$ és $\mathbf{p}_{3D} = [X \quad Y \quad Z \quad 1]^T$ pontokat a 2×4-es $\mathbf{B}$ affin mátrix kapcsolja össze: $\mathbf{p}_{2D} = \mathbf{B}\mathbf{p}_{3D}$.



2. ábra. A javasolt affín becslők kvantitatív összehasonlítása szintetikus bemenet esetén. Általános stereo konfigurációt veszünk figyelembe. A zajt a sakktábla celláinak 2D-s irányaihoz nulla várható értékű Gauss-eloszlással adtam hozzá, amelynek szórása $3,0^\circ$ volt.

de algebrailag nem egyszerűsödik a normálvektorral közös összefüggés, ezért algoritmikus haszna nincsen az affin kamerának.

4. Kérdések a 4. fejezethez

A minimal solver-eknél a degeneráció és numerikus instabilitás kulcskérdés. Melyek a szerző által legfontosabbnak tartott degenerált konfigurációk az affin-alapú solver-ek esetén?

A bíráló ezzel a kéréssel is megfogott, egyrészt mert ezt nem vizsgáltuk meg alaposan, másrészről pedig azért, mert a disszertációban tíznél több megoldó van, és a kérdést mindegyikre külön-külön kell megvizsgálni. Ha sikerrel járunk, akkor akár egy önálló szakcikket is lehet minden egyes témáról írna. Például én is csak az idén olvastam róla, hogy a fundamentális mátrix hagyományos, pont alapú becslésénél a kézenfekvő eseten kívül (a pontok egy síkban vannak) egy másik degeneráció is lehetséges: ha van egy hiperboloid/kúp/henger, amely a 3D pontokat és kamera központokat tartalmazza [LF96]. Ezt is onnan tudom, hogy bíráltam a CVPR2026-os konferenciára egy kéziratot, amely még 2026-ban is azzal foglalkozik, hogyan lehet ezeket a degenerált eseteken hatékonyan detektálni.

De nem szeretnék a kérdés elől teljesen kitérni, ezért gondolkoztam egy kicsit a problémán, és – egyelőre publikálatlan – levezetéseket is végeztem. Azt tartottam célravezetőnek, ha az affin transzformációkat felírom a kamerák külső paramétereinek függvényeként, hátha beszédesebb lesz, mint a korábban publikált alakok. Szerencsémre nem csalódtam az eredményekben.

Ahogy megvizsgáltam a problémát, arra jutottam, hogy az instabilitások felderítéséhez érdemes lenne egy új levezetést megalkotni. Ha normalizált koordinátákat alkalmazunk, azaz a

$$\mathbf{K}_1^{-1}\begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix},$$
$$\mathbf{K}_2{}^{-1}\begin{bmatrix} u_2 \\ v_2 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} u_2 \\ v_2 \\ 1 \end{bmatrix}$$

helyettesítéseket alkalmazzuk, akkor a kamera belső paramétereitől meg tudjuk tisztítani az összefüggéseket. Ez a normalizálás a gyakorlatban is elvégezhető, ha a kameránk kalibrált, azaz a belső paramétereket tartalmazó $\mathbf{K}_1$ és $\mathbf{K}_2$ mátrixok ismertek. Ha az érintősíkot implicit alakban leírva, azaz a $n_x x + n_y y + n_z z + d = 0$ egyenletet használva írjuk fel, ahol $\mathbf{n} = [n_x \quad n_y \quad n_z]^T$ a sík normálisa és $[x \quad y \quad z]^T$ pedig egy tetszőleges pontja, akkor a két kép között csak az $\mathbf{R}$ elforgatás és a $\mathbf{t}$ eltolás választható meg szabadon.

Ahogyan azt a függelékben levezetem, az affin transzformációt kalibrált kamerák esetében az alábbi alakra lehet hozni:

$$\mathbf{A} = \frac{1}{s}\left(\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} - \begin{bmatrix} u_2 \\ v_2 \end{bmatrix}\begin{bmatrix} R_{31} & R_{32} \end{bmatrix} - \frac{1}{d}\begin{bmatrix} t_x - u_2 t_z \\ t_y - v_2 t_z \end{bmatrix}\begin{bmatrix} n_x & n_y \end{bmatrix}\right), \tag{1}$$

ahol d a sík és az első kamera fókuszpontjának a távolsága, $s = \left(\mathbf{r}_3^T + \frac{t_z}{d}\mathbf{n}^T\right)\mathbf{p}_1$, amennyiben $\mathbf{r}_3^T$ az $\mathbf{R}$ forgatási mátrix harmadik sora és $\mathbf{p}_1 = [u_1 \quad v_1 \quad 1]^T$ az első pont homogén koordinátás alakja.

Ha a degenerált eseteket tekintjük, akkor a nullával való osztás lehetőségét meg kell vizsgálnunk. Ez két esetben fordulhat elő: $d = 0$ és $s = 0$.

- Az első esetben az érintősík átmegy az első kamerán, hiszen nulla a távolság. Ez akkor fordulhat elő, ha pont a sík "élét" látjuk az első kamerából. Azaz a rajta lévő mintázat végtelenül kicsi. Ebben az esetben a képekből már az affin transzformációt se lehet kinyerni, hiszen a minta nem látszik.

- A második triviális eset az $s = 0$ pedig akkor fordulhat elő matematikailag, ha az $\mathbf{r}_3 + \frac{t_z}{d}\mathbf{n}$ vektor merőleges az első lépen levő pont homogén koordinátás $\mathbf{p}_1$ alakjára. Ez geometriailag annak a ténynek felel meg, és általános homográfiára is igaz [HZ03], hogy a megvizsgált térbeli hely mélysége a második kamerából nézve pontosan nulla. Ilyen a valóságban nem fordulhat elő, mert akkor az a rész nem látszódik a kamerán.

Amennyiben $d \neq 0$ és $s \neq 0$, az affin transzformáció mindig értelmezhető, még az is csak ritka esetben fordulhat elő, hogy az értéke nulla.

Azonban az jól látszik, hogy az affin összefüggés jobb oldalán szerepel a tag:

$$\frac{1}{d}\begin{bmatrix} t_x - u_2 t_z \\ t_y - v_2 t_z \end{bmatrix}\begin{bmatrix} n_x & n_y \end{bmatrix}. \tag{2}$$

Csak ebben szerepel a normálvektor, illetve még a közös osztóban: $s = \left(\mathbf{r}_3^T + \frac{t_z}{d}\mathbf{n}^T\right)\mathbf{p}_1$.

A normálvektor becslése esetén egy olyan $\mathbf{n} = [n_x \quad n_y \quad n_z]^T$ vektort keresünk, amelynek egységnyi a hossza. A három koordinátára így kettő szabadságfok jut, tehát legalább kettő egyenletünk van. Viszont, ha fennáll, hogy

$$\begin{aligned} t_x - u_2 t_z &= 0, \\ t_y - v_2 t_z &= 0, \end{aligned} \tag{3}$$

akkor a fenti tag kiesik, és ezzel n_x és n_y eltűnik az affin összefüggésből. Azaz nem lesz elég információnk ahhoz, hogy a normálvektort megbecsüljük, mert csak a teljes s értékét lehet becsülni, azaz $\mathbf{n}^T\mathbf{p}$-re lesz egy skalár értékünk. Ezért $\mathbf{n}$-et nem tudjuk kinyerni, csak annyit tudunk, hogy rajta lesz egy köríven, és ennek a körlapnak a normálvektora párhuzamos lesz $\mathbf{p}$-vel.

A bemutatott gyorsulás milyen arányban köszönhető a kisebb mintaszámnak, és milyen arányban annak, hogy az affin információ miatt kevesebb „téves" hipotézis jut át a robusztus kiértékelésen?

Első megfontolás után egyértelműen azt válaszolnám, hogy mivel a RANSAC algoritmus futási ideje nem polinomiális összefüggésben növekszik a minimális minták számával, ezért a futási idő emiatt csökkenthető, nagy mintaszámcsökkentés esetén akár radikálisan is.

Azonban azt is meg kell jegyeznem, hogy az affin transzformációból jövő paraméterek információtartalmának mérésével tudtommal eddig nem foglalkoztak. Magam a disszertáció benyújtása óta igyekeztem a becslési feladatok kondicionáltságával foglalkozni, miután négy affin paramétert is fel tudunk használni, de nem mindegy, hogy melyiket milyen súllyal. Erre a kondicionáltság egy jó mérőszám lehet, ami az információtartalommal is összefügg.

Egyszerűbb esetekre vannak ugyan kezdeti eredményeim, de a megfontolásaim még csak alapvető összefüggéseken alapulnak, ezért válasz helyett ezt a részt a jövő feladatai közé tenném, és egyben megköszönném a Bírálónak, hogy rámutatott a kérdés fontosságára.

5. Kérdések az 5. fejezethez:

Itt sajnos bevezetésként meg kell jegyeznem, hogy a fotometrikus sztereóval kapcsolat munkákat több, mint tíz éve végeztük Fodor Bálinttal, akivel már a kapcsolat megszakadt. Így a módszerek újrafuttatásához sajnos mindent elölről le kellene futtatni, már a tesztelési adatok sincsenek meg.

1.2.Fotometrikus sztereó esetén a reflektancia és megvilágítás modellezése kritikus. Mennyire érzékeny a bemutatott optimalizációs keret a Lambert-model sérülésére (specularitás, változó albedo)?

A visszaverődés modellezésére a Lambert-törvényt alkalmaztuk, amely ideális matt felületet feltételez. Ennél a modellnél a visszavert fény intenzitása a megvilágítás erősségétől, az albedótól, valamint a fényirány és a felületi normális által bezárt szög koszinuszától függ. Fontos tulajdonsága, hogy a felület a fényt minden irányba egyformán szórja szét (izotróp), így a megfigyelt fényesség nem függ a nézőponttól.

A bíráló felvetése helyes, hogy a spekuláris visszaverődés hibát okoz. Ebben az esetben a fény a tükörirányba terjed, vagy annak közelébe. Ilyen vizsgálatokat nem végeztünk, a teszteléseknél nagyon figyeltünk arra, hogy ne csillogó anyagokat használjunk. Ehhez a lézeres szkennelésnél is rendszeres használt fehér por alkalmazására is szükség volt.

A matt és a spekuláris felületek annyira különbözőképpen viselkednek, hogy biztosra vehető, hogy a kidolgozott módszerünk nem fog a csillogó felületre működni.

További problémát jelent, hogy mivel a csillogás a tükörirányban jelentős, nem elég, ha egy pozícióban van a kamera, hanem célszerű lenne a kamerát (vagy a tárgyat) mozgatni. Tehát sok fényirányt kellene alkalmazni, sok kamerapozícióhoz. Így a probléma sokkal bonyolultabb lenne, de a Jakobi mátrix ritkasága továbbra is fennállna.

Az irányított pontfelhő rekonstrukció esetében mekkora szerepe van a kezdeti becslésnek: mennyire nagy a konvergenciatartomány, illetve milyen tipikus rossz lokális optimumok jelenhetnek meg?

A kötegelt behangolás algoritmus esetén (SfM - structure from motion feladatra) a konvergencia kérdését már elég sokan vizsgálták (pl. [ESN06, EBCI16]). Sajnos a rövid válasz az, hogy elég szűk a tartomány, tele van lokális minimumokkal az optimalizálandó függvény.

Ezért egy jó kezdeti becslés szükséges a numerikus minimalizáláshoz. Ez a visszavetítési helyekre maximum néhányszor tíz pixeles pontosságot jelent, az affin transzformáció esetén még két nagyságrenddel pontosabbnak kell lenni. De a célfüggvény érzékeny sok mindenre, a felületi normálisoktól a kamerák bázistávolságáig.

Ez a terület is megérne egy újabb kutatási munkát, mert az irányított pontfelhőre még senki nem végzett ilyen munkát. Erre a munkánkra eddig öt független hivatkozás érkezett, ezért az gyanítható, hogy nem sokan próbálták reprodukálni.

6. Kérdések az 6. fejezethez:

A szerző tapasztalata szerint mi dominálja leginkább a végső kalibrációs hibát: a célobjektum detektálása (képen/pontfelhőben), a szenzorok időszinkronizációja, vagy a célobjektum geometriai modellhibája?

A kamerák időben össze vannak szinkronizálva, ezért közöttük nincsen időbeli elcsúszás. De eleve figyelünk arra, hogy akár a doboz, akár a gömb a felvételeken ne mozogjon, ezért ez okozza a legkisebb, szinte elhanyagolható hibát.

A mi LiDARunk (Velodyne VLP-16) mérési pontossága egy centiméteres nagyságrendben van, de elég ritka a pontfelhő, ezért ha kicsi a kalibrációs tárgy, az azért probléma, mert kevés pontot fog szkennelni a felületből, és kevesebb pontból rosszabb becslést kapunk. A gömbünk sugara $30cm$, a dobozok oldalainak mérete is ebben a nagyságrendben van. Nagyobb gömb/doboz érezhetően jobb kalibrációs eredményt adna.

De a legnagyobb hiba egyértelműen a doboz alakjából fakad, a kartondobozok hajlékonyak, sőt sérülékenyek. (Cserébe viszont könnyű beszerezni.) A gömb hungarocellből készült, nagyon pontos, ezért az alak esetleges torzulása elhanyagolható hibát okoz.

A célobjektum-alapú kalibrációk kontrollált mérést igényelnek. A szerző lát-e reális átmenetet target nélküli („targetless") LiDAR–kamera kalibráció felé a saját módszerei szellemiségében?

Mi szeretjük online kalibrációnak hívni, amikor menet közben kell az eszközök közötti külső paramétereket ('póz', azaz forgatás+eltolás) hangolni. Ez egy rendkívül népszerű terület, Dunát lehet rekeszteni a módszerekkel. Egy viszonylag friss áttekintő cikk ezen a referencián keresztül elérhető: [ADQ+24]

A mi módszereink kalibrációs objektumai közül a szabályos gömb valós helyzetben ritkán fordul elő, ezért csak a dobozos megoldás képzelhető el, ahol egymásra merőleges síkokat feltételezünk. Az ember alkotta világban merőleges felületek elég gyakran adódnak. Például egy utcasarkon az épületek két fala és a talaj három merőleges síkot alkot, ezért itt elképzelhetőnek tartanám a dobozos kalibrációs eljárást kisebb módosításokkal, hiszen a felületek bár merőlegesek, de nem ugyanolyan iránnyal (normálvektorokra tekintettel) rendelkeznek. A pontfelhőn merőleges síkok hatékony szegmentálása megoldott feladatnak tekinthető, a nagyobb nehézség a kameraképeken a síkfelületek automatikus detektálása. Ehhez 2026-ban vagy monodepth [LCL+25] (egyképes rekonstrukciós) vagy valamilyen hagyományos, RANSAC-alapú, homográfia-illesztésen keresztül történő síkbecslést javasolnék, például a II.4. tézispont alapján.

Megjegyezném, hogy a IV.3. tézisben ismertetett módszer előnye, hogy ugyan nem 'targetless', de egy falra rögzített sakktábla esetén, például a garázsban, alkalmas a kalibráció folyamatos ellenőrzésére.

Véleményem szerint ugyanakkor a 'targetless' kalibrálás legnagyobb előnye, hogy nem egy tárgyhoz vagyunk kötve, és időben is nagy a szabadság, azaz folyamatosan lehet adatokat gyűjteni a kalibrációhoz. Az én téziseim adott geometriához kötöttek, az időbeliség ezért csak olyan értelemben használható, hogy a kalibrációs tárgyat (esetleg tárgyakat) különböző nézőpontból lehet megtekinteni.

Kelt: Budapest, 2026. március 13.

Hajder Levente

Hivatkozások

[ADQ+24] Pei An, Junfeng Ding, Siwen Quan, Jiaqi Yang, You Yang, Qiong Liu, and ma Jie. Survey of extrinsic calibration on lidar-camera system for intelligent vehicle: Challenges, approaches,

and trends. *IEEE Transactions on Intelligent Transportation Systems*, PP:1–25, 11 2024.

[BM04] Simon Baker and Iain A. Matthews. Lucas-kanade 20 years on: A unifying framework. *Int. J. Comput. Vis.*, 56(3):221–255, 2004.

[EBCI16] Anders P. Eriksson, John Bastian, Tat-Jun Chin, and Mats Isaksson. A consensus-based framework for distributed bundle adjustment. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1754–1762, 2016.

[ESN06] Chris Engels, Henrik Stewénius, and David Nistér. Bundle adjustment rules. In *Photogrammetric Computer Vision*, 01 2006.

[FL88] Olivier Faugeras and F. Lustman. Motion and structure from motion in a piecewise planar environment. Technical Report RR-0856, INRIA, 1988.

[Haj26a] Levente Hajder. Affine correspondences in stereo vision: Theory, practice, and limitations, 2026. https://arxiv.org/abs/2603.01836.

[Haj26b] Levente Hajder. A unified formula for affine transformations between calibrated cameras, 2026. https://arxiv.org/abs/2602.06805.

[HZ03] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[LCL+25] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[LF96] Quang-Tuan Luong and Olivier D. Faugeras. The fundamental matrix: Theory, algorithms, and stability analysis. *Int. J. Comput. Vis.*, 17(1):43–75, 1996.

[LK81] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In Patrick J. Hayes, editor, *Proceedings of the 7th International Joint Conference on Artificial Intelligence, IJCAI '81, Vancouver, BC, Canada, August 24-28, 1981*, pages 674–679, 1981.

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.

[RRKB11] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. ORB: an efficient alternative to SIFT or SURF. In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool, editors, *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 2564–2571. IEEE Computer Society, 2011.

[SGY+25] Pengju Sun, Banglei Guan, Zhenbao Yu, Yang Shang, Qifeng Yu, and Daniel Barath. Learning affine correspondences by integrating geometric constraints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 27038–27048. Computer Vision Foundation / IEEE, 2025.

[WLC+24] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, June 2024.

[YKH+24] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 10371–10381. Computer Vision Foundation / IEEE, 2024.

A. Az affin transzformációk alakja kalibrált kamerák esetén

Ahogyan az a sztereó látásban ismert [FL88], egy $\mathbf{H}$ homográfia kalibrált kamerák esetén felírható ebben az alakban:

$$\mathbf{H} = \mathbf{R} - \frac{\mathbf{t}\mathbf{n}^T}{d} = \begin{bmatrix} R_{11} - \frac{t_x n_x}{d} & R_{12} - \frac{t_x n_y}{d} & R_{13} - \frac{t_x n_z}{d} \\ R_{21} - \frac{t_y n_x}{d} & R_{22} - \frac{t_y n_y}{d} & R_{23} - \frac{t_y n_z}{d} \\ R_{31} - \frac{t_z n_x}{d} & R_{32} - \frac{t_z n_y}{d} & R_{33} - \frac{t_z n_z}{d} \end{bmatrix},$$

ahol az $\mathbf{R}$ ortonormált mátrix és a $\mathbf{t} = [t_x \quad t_y \quad t_z]^T$ vektor adja a két kamera közötti forgatást és eltolást (együtt: 'póz'), továbbá $\mathbf{n} = [n_x \quad n_y \quad n_z]^T$ a sík normálvektora, d pedig a sík távolsága az első kamerától. Mindez, behelyettesítve a mátrix és a vektorok elemeit adja az alábbi összefüggést:

$$\mathbf{H} = \begin{bmatrix} R_{11} - \frac{t_x n_x}{d} & R_{12} - \frac{t_x n_y}{d} & R_{13} - \frac{t_x n_z}{d} \\ R_{21} - \frac{t_y n_x}{d} & R_{22} - \frac{t_y n_y}{d} & R_{23} - \frac{t_y n_z}{d} \\ R_{31} - \frac{t_z n_x}{d} & R_{32} - \frac{t_z n_y}{d} & R_{33} - \frac{t_z n_z}{d} \end{bmatrix}.$$

Az affin transzformáció a disszertációban második törvényként megadott összefüggések [6] alapján az alábbiak szerint írható fel:

$$a_{11} = \frac{H_{11} - H_{31}u_2}{(H_{31}u_1 + H_{32}v_1 + H_{33})} = \frac{b_{11}}{s}.$$

Vezessük be az alábbi jelöléseket:

$$b_{11} = R_{11} - \frac{t_x n_x}{d} - u_2\left(R_{31} - \frac{t_z n_x}{d}\right) = R_{11} - u_2 R_{31} - \frac{n_x}{d}\left(t_x - u_2 t_z\right)$$

és

$$s = H_{31}u_1 + H_{32}v_1 + H_{33} = \left(R_{31} - \frac{t_z n_x}{d}\right)u_1 + \left(R_{32} - \frac{t_z n_z}{d}\right)v_1 + \left(R_{33} - \frac{t_z n_y}{d}\right) = \mathbf{r}_3^T\mathbf{p} + \frac{t_z}{d}\mathbf{n}^T\mathbf{p} = \left(\mathbf{r}_3^T + \frac{t_z}{d}\mathbf{n}^T\right)\mathbf{p},$$

ahol a $\mathbf{p}_1$ vektor tartalmazza az első képen a megfeleltetett pont koordinátáit homogén alakban, azaz $\mathbf{p} = [u \quad v \quad 1]^T$, $\mathbf{r}_3$ vektor pedig az $\mathbf{R}$ forgatási mátrix harmadik sorát jelöli.

Hasonlóan,

$$a_{12} = \frac{b_{12}}{s}, \quad a_{21} = \frac{b_{21}}{s}, \quad a_{22} = \frac{b_{22}}{s},$$

amennyiben

$$b_{12} = h_{12} - h_{32}u_2 = R_{12} - \frac{t_x n_y}{d} - u_2\left(R_{32} - \frac{t_z n_y}{d}\right) =$$

$$R_{12} - u_2 R_{32} - \frac{n_y}{d}\left(t_x - u_2 t_z\right),$$

és

$$b_{21} = h_{21} - h_{31}v_2 = R_{21} - \frac{t_y n_x}{d} - v_2\left(R_{31} - \frac{t_z n_x}{d}\right) = R_{21} - v_2 R_{31} - \frac{n_x}{d}\left(t_y - v_2 t_z\right),$$

[6] 16-os formulák a disszertáció 21. oldalán

végezetül pedig

$$b_{22} = h_{22} - h_{32}v_2 = R_{22} - \frac{t_y n_y}{d} - v_2\left(R_{32} - \frac{t_z n_y}{d}\right) = R_{22} - v_2 R_{32} - \frac{n_y}{d}\left(t_y - v_2 t_z\right).$$

Ezeket visszahelyettesítve a teljes affin transzformációra adódik, hogy

$$\mathbf{A} = \frac{1}{s}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \frac{1}{s}\begin{bmatrix} R_{11} - u_2 R_{31} - \frac{n_x}{d}\left(t_x - u_2 t_z\right) & R_{12} - u_2 R_{32} - \frac{n_y}{d}\left(t_x - u_2 t_z\right) \\ R_{21} - v_2 R_{31} - \frac{n_x}{d}\left(t_y - v_2 t_z\right) & R_{22} - v_2 R_{32} - \frac{n_y}{d}\left(t_y - v' t_z\right) \end{bmatrix} = \frac{1}{s}\left(\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} - \begin{bmatrix} u_2 \\ v_2 \end{bmatrix}\begin{bmatrix} R_{31} & R_{32} \end{bmatrix} - \frac{1}{d}\begin{bmatrix} t_x - u_2 t_z \\ t_y - v_2 t_z \end{bmatrix}\begin{bmatrix} n_x & n_y \end{bmatrix}\right)$$

Ebben az alakban az affin transzformáció három 2×2-es mátrix összegére bontható, amelyikből az első a forgatási mátrix bal felső blokkja, a második és a harmadik elem egy-egy diád. Érdekesség, hogy a harmadik tartalmazza a normálvektor első két koordinátáját, de ez megtévesztő lehet, mert az s osztó is függvénye a normálvektor mindhárom elemének.