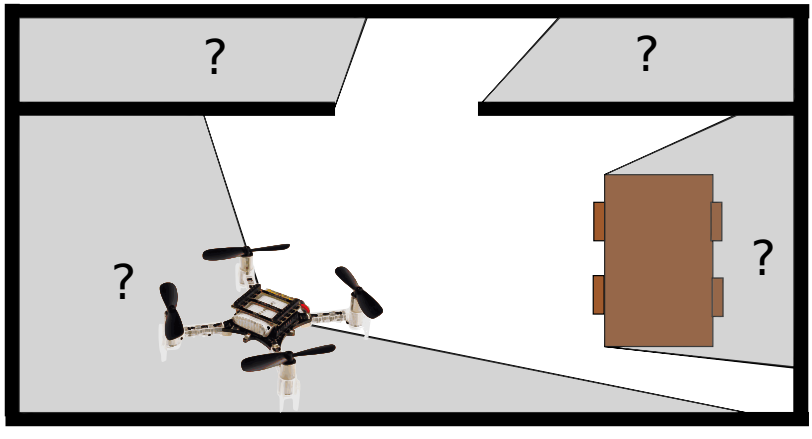
Populärvetenskaplig sammanfattning

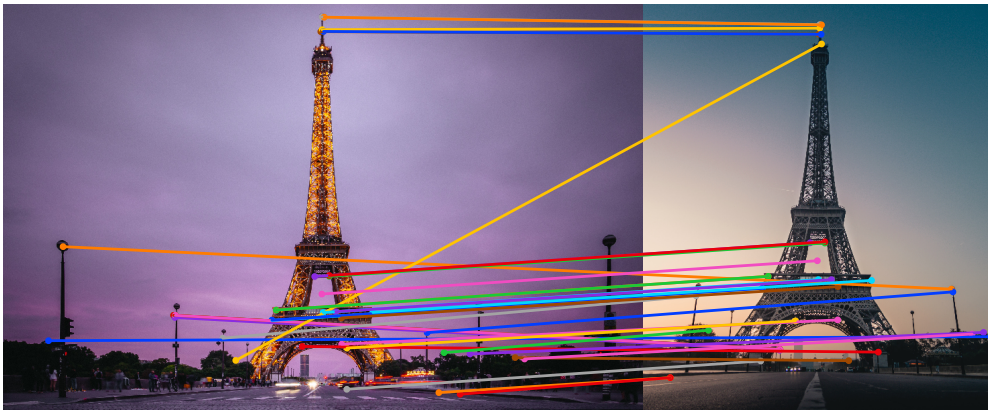
Som människor har vi förmågan att uppfatta alla tre spatiella dimensioner (3D) i vår omgivning, enbart med hjälp av våra ögon samt förflyttning av kroppen genom denna omgivning. Trots att våra ögon – liksom kameror – enbart ser en platt projektion av världen, så har vi lärt oss att tänka på den i 3D. Till exempel, genom att gå runt i ett hus lär vi oss hur rummen är placerade och dess storlek, genom att promenera runt ett kvarter kan vi få en känsla för var vi är, med hjälp av erfarenhet kan vi uppskatta höjden på ett skåp utan att ta fram måttbandet.

Kan vi lära en dator att göra detta? Detta arbete studerar närliggande frågor inom forskningsfältet *datorseende*. Som verktyg används geometriska matematiska modeller tillsammans med *neurala nätverk*, som är komplexa matematiska modeller inspirerade av hjärnans anatomi och behöver stora mängder data att lära sig ifrån. Exempel på frågor som studeras är: Kan vi träna ett neuralt nätverk att resonera kring dolda ytor inomhus som i Figur 1? Kan vi träna neurala nätverk att lista ut ett rums utformning i termer av golv, väggar och tak från en bild? Hur kan vi på ett effektivt sätt utnyttja rörelse för att skatta en modell i 3D från bilder tagna på olika platser?



Figur 1: En drönare i ett rum sett ovanifrån. Den försöker gissa vad som kan finnas dolt bakom väggarna och bordet.

Hjärnans förmåga till abstrakt tänkande är en av många faktorer som bidrar till vår orienteringsförmåga. För att orientera oss kan vi notera intressanta objekt – hus, skyltar, träd m.m. – som hjälper oss att förstå hur vi rör oss och ger oss möjlighet att hitta bättre när vi återbesöker samma plats. Inom datorseende finns det ett relaterat problem som kallas *Structure from Motion* (SfM), d.v.s. struktur från rörelse. Målet är att från bilder skapa en 3D-modell av omgivningen genom att beräkna kameras position och 3D-modellen samtidigt. Det första steget är typiskt sett att hitta intressanta objekt eller punkter i bilderna som kan användas för att relatera bilderna till varandra.

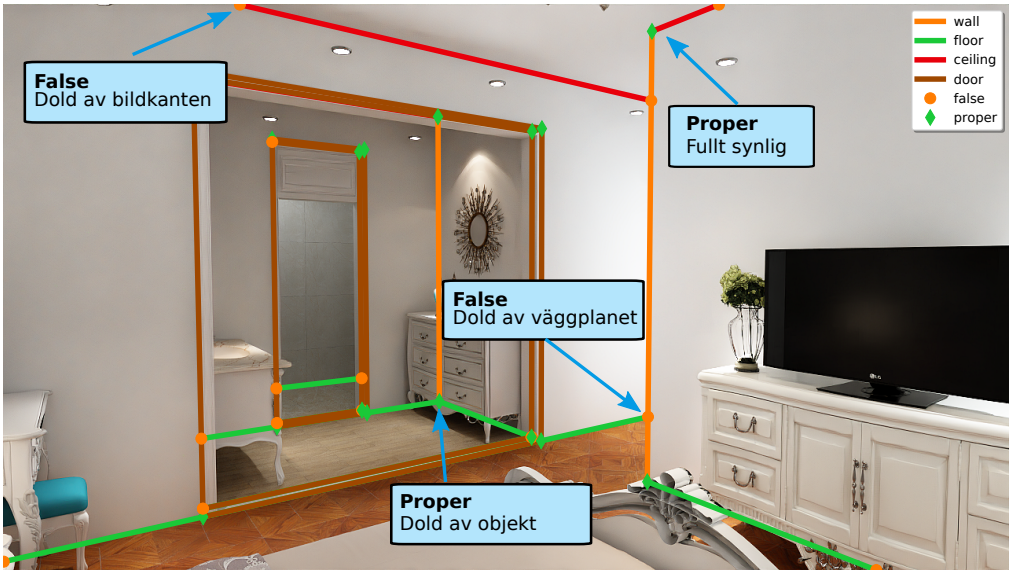


Figur 2: Två bilder tagna på Eiffeltornet vid olika tidpunkter och på olika platser. Intressepunkter (SIFT) har hittats i varje bild och en matchning baserad på utseendet har gjorts. Vi ser att vissa punkter matchar bra men delar av tornet saknar matchningar och ett flertal är helt felaktiga.
Bildkälla: Denys Nevozhai och Gautier Salles från `unsplash.com`.

Standardlösningen är att använda *intressepunkter* i bilden, med en unik färgsignatur som kallas *deskriptor*. Denna deskriptor används för att hitta matchande punkter i de andra bilderna. Detta är oftast en bra lösning men den har svårt att hantera upprepande mönster som tegelväggar eller blanka ytor som t.ex. vitmålade väggar. Förändringar över tid är också svårt, till exempel så ser ett träd på vintern väldigt annorlunda ut jämfört med ett träd på sommaren i Sverige. I Figur 2 ser vi automatiskt detekterade intressepunkter som är matchade mellan två bilder på Eiffeltornet, tagna vid olika tidpunkter. Under dessa förhållanden går det enbart att matcha vissa delar av tornet och många matchningar är felaktiga.

För att hantera dessa problem studeras i denna avhandling hur t.ex. träd och stolpar kan modelleras som parallella cylindrar för att förbättra robusthet och effektivitet. Dessutom studeras hur linjer och polygoner kan detekteras automatiskt och användas för att representera ett rums utformning i termer av golv, väggar, tak, fönster och dörrar som illustreras i Figur 3. Dessa objekt har en högre abstraktionsnivå än intressepunkter och kan förbättra robustheten i SfM, vilket är nödvändigt för att nå samma nivå av 3D-förståelse för en dator som för en människa.

När intressepunkter eller objekt har hittats är det dags att skatta position och orientering både av objekten samt av kamerorna som användes för att ta bilderna. För att hitta korrekta matchningar av objekt samt punkter mellan par av bilder behövs en robust matchningsmetod som kan ignorera de dåliga matchningarna. En vanlig metod för detta är *RANdom SAmple Consensus* (RANSAC), som löser ett *minimalt problem* för en mängd av slumpmässigt valda punkter, om och om igen. Det är viktigt att utveckla snabba *minimallösare* för dessa minimala problem, eftersom det möjliggör matchning av bilderna i realtid och i stor skala. Till exempel, för att skatta relativ position och orientering mellan två kalibrerade



Figur 3: Artikel II föreslår en wireframe-representation för rum. Linjerna och punkterna kan användas för SfM och är stabila eftersom de beror på rummets utformning snarare än dess utseende gällande t.ex. möbler och tapeter.

kameror så krävs minst fem matchande punkter i varje bild. Problemet har studerats av flertal forskare och det finns minimallösare som löser problemet på mikrosekunder. Detta innebär att vi kan köra tusentals iterationer i RANSAC för att hitta den bästa mängden punkter för matchning – på bara några millisekunder. I denna avhandling presenteras minimallösare för både matchning av parallella cylindrar samt flexibel matchning av intressepunkter.

När matchningen är gjord behöver position och orientering för alla punkter och kameror optimeras för att minimera felen i 3D-modellen, som vi också benämner *kartan*. Detta kallas för *bundle adjustment* och är en iterativ optimering som minskar *återprojiceringsfelet*, vilket är felet mellan den observerade intressepunkten och den skattade 3D-punkten projicerad tillbaka till bilden. Om enbart intressepunkter används så är den färdiga kartan en mängd av punkter i 3D – vilket kallas *punktmoln* – som representerar strukturen i omgivningen.

Att utföra bundle adjustment på ett storskaligt problem är svårt eftersom de nödvändiga beräkningarna inte skalar proportionerligt mot antalet bilder. Detta innebär att även om vi med en kan dator processa utsidan av en byggnad på några minuter skulle det ta dagar att modellera hela kvarteret med samma metod. Det finns förstås sätt att göra det möjligt, oftast genom att dela upp problemet i mindre bitar. I denna avhandling studeras *map merging*, där frågan är hur vi på ett effektivt sätt kan slå samman två eller flera punktmoln utan att göra nya bundle adjustments? Metoder presenteras som optimerar återprojiceringsfelet

utan att kräva fullständigt överlapp av punktmolnen. De reducerar nödvändiga beräkningar men är fortfarande flexibla nog att korrigera formen på de ingående punkmolnen.

Sammanfattningsvis studeras flera aspekter av SfM i denna avhandling, från detektion av linjer och polygoner till minimallösare och sammanslagning av punktmoln. Förhoppningsvis kan dessa bidrag vara pusselbitar som bidrar till att autonoma system och andra tjänster kan förstå vår omgivning på samma sätt som vi gör.