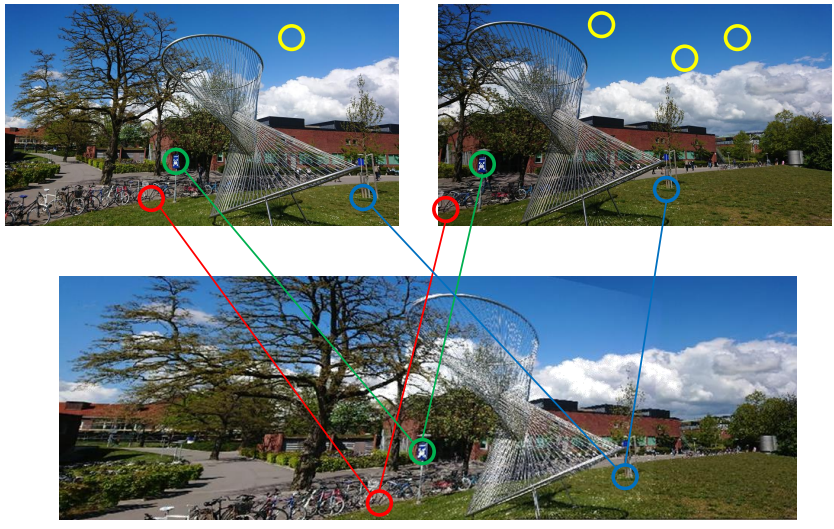
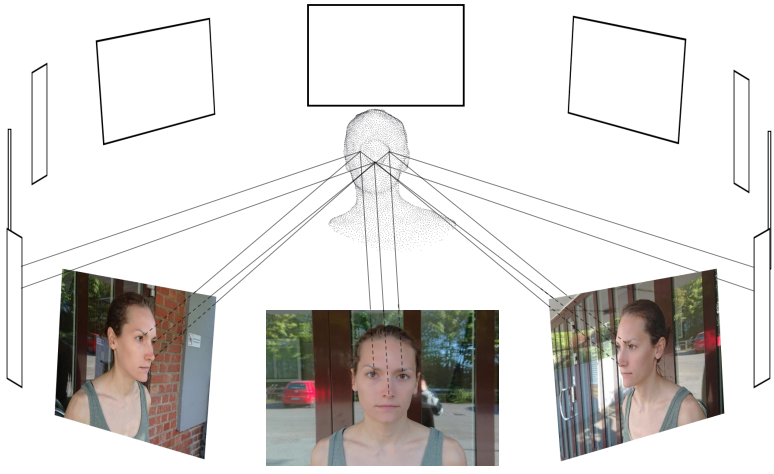
Populärvetenskaplig sammanfattning

Kan datorer förstå världen lika bra som människor? Kan en robot navigera på egen hand? Kan en drönare komma ihåg hur ett rum ser ut? Och kan man från platta bilder förstå hur omgivningen de avbildar ser ut? Allt detta är frågor med koppling till innehållet i denna avhandling.



Bilden visar sammansättning av två olika bilder som fångar samma scen till en panoramabild. För detta krävs att man identifierar intressanta och matchande punkter i de båda bilderna, exempelvis de som är markerade av de röda, gröna och blå cirklarna. De gula cirklarna markerar punkter som är dåliga att använda, eftersom dessa inte är unika.

Datorseende är ett område som handlar om att lära datorer att förstå och utläsa information ur digitala bilder, precis som vi människor gör när vi ser någonting. Det mänskliga ögat är i många avseenden likt en kamera och generellt är människor och djur väldigt bra på att förstå saker – som vad det är man ser eller hur långt bort olika saker är. Datorseende handlar om att lära datorer att göra samma sak. En sak som bilder kan användas till är att skapa *3D-modeller* eller *kartor* av verkligheten. Vi människor kan bedöma djupet i det vi ser, och med hjälp av detta, vårt minne och våra erfarenheter kan vi skapa oss våra egna kartor av hur exempelvis ett rum eller en lägenhet ser ut. Samma sak kan man göra digitalt, med hjälp av datorer. Tillvägagångssättet är ganska likt det man använder när man skapar panoramabilder. I dag har de flesta mobiltelefoner både en kamera och en funktion för att skapa panoramabilder – det vill säga stora bilder som egentligen är flera ihopklistrade bilder. Om man vill klistra ihop två bilder gör man väsentligen så att man noterar *intressanta* punkter som syns i båda bilderna och ser till att dessa matchar. Ett exempel på detta syns i figuren ovan.

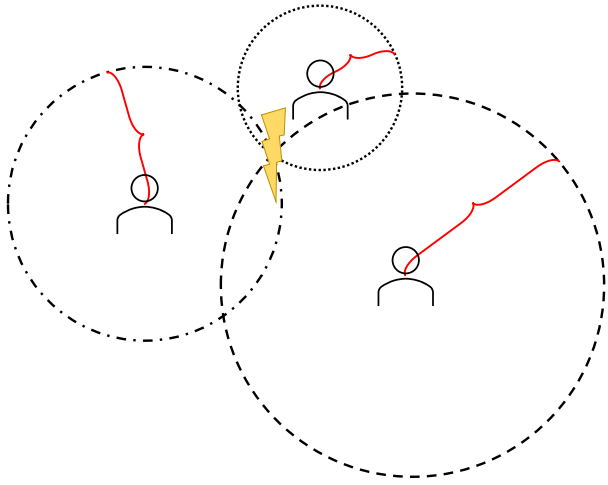


Med hjälp av flera bilder bestäms positionen i 3D för de ”intressanta” punkterna. Om man har tillräckligt många bilder kan man skapa en 3D-modell, här i form av ett moln av punkter.

På samma sätt kan man, om man har tillräckligt många bilder från olika vinklar, “klistra ihop” dem till en 3D-modell. Djupet fås, precis som när människor ser, av att bilderna är tagna från olika ställen, se bilden ovan. Att med hjälp av kamerapositioner och intressanta punkter i bilder beräknar positionen för 3D-punkter kallas det för *triangulering* och om man dessutom samtidigt hittar kamerapositionerna brukar det kallas *struktur och rörelse*, eller *structure from motion*. När tillräckligt många punkter har triangulerats har man en typ av karta av omgivningen.

Sådana här kartor kan även skapas med hjälp av andra sensorer och signaler, till exempel mikrofoner och ljud. Precis som att en bild beskriver någon form av information i två dimensioner så gör ljudet det i en dimension. De allra flesta har nog vid något tillfälle räknat sekunderna från det att man ser en blixt tills dess att man hör åskknallen. Därigenom vet man hur långt bort åskan är. Man vet inte vilket håll den kommer ifrån, men man vet att den befinner sig någonstans på en cirkel där man själv är centrum och det uträknade avståndet är radien. Om man dessutom känner till vad avståndet till några fler punkter är kan man rita upp fler cirklar och där dessa skär varandra befinner sig åskan för stunden. Detta finns illustrerat i figuren på nästa sida. Samma sak kan man göra med vanliga mikrofoner och högtalare som står uppställda i ett rum. Finns det tillräckligt många så räcker det att man vet de respektive avstånden för att kunna beräkna de relativa positionerna både för mikrofonerna och för högtalarna. Därigenom får man en karta över mikrofonpositionerna, som på många sätt är lik den karta man kan skapa med hjälp av bilder.

När man väl har 3D-kartor kan dessa användas för *positionering*, alltså för att ta reda på var



Bilden illustrerar hur man kan lista ut var exempelvis ett åskoväder befinner sig, om man är tre personer som hör åskan från olika platser. Genom att rita cirklar runt de olika personerna med radier som motsvaras av avståndet, så vet man att åskan är där cirklarna skär varandra.

man befinner sig, antingen med hjälp av ljud eller bilder. Ett exempel på ett system som använder signaler som liknar ljud är GPS, som används flitigt för utomhuspositionering och navigation. Inomhus fungerar dock GPS sämre, och därför kan det vara bra att ha andra system som kan användas på liknande sätt men för inomhuspositionering. För att positioneringen ska bli så exakt som möjligt är det viktigt att kartan är så exakt som möjligt och kartorna blir generellt bättre ju fler mätningar man gör (exempelvis ju fler bilder man använder). Därför är det bra om man kan slå samman flera olika kartor av samma miljö, för att öka noggrannheten och för att kunna uppdatera kartan om någonting ändras. Detta kan man göra genom att – precis som i bilderna – identifiera motsvarande intressanta punkter i de olika kartorna och pussla ihop dem så att de överlappar.

Ett exempel på när kartsammanslagning kan vara användbart är för självkörande bilar. Många bilar har i dag kameror och dessa kan användas för att bestämma positionen för bilen i en sedan tidigare känd miljö. Vi tänker oss att vi har en karta över en stad och att en bil sedan kör igenom denna stad. Medan detta händer kommer bilen samla in en massa bilder och den kan då skapa sig sin egen, *lokala* karta av de delar av staden som den passerar. Om denna lokala karta sedan kan läggas till den stora, *globala* kartan, så kommer den därefter att innehålla mer information och därmed vara med exakt. Om alla bilar som kör genom staden kan göra detta kommer kartan successivt att bli bättre och om någon infrastruktur i staden ändras kommer detta att återspeglas i kartan utan att den helt behöver göras om.

I den här avhandlingen fokuserar vi först på att hitta exakta mätningar för avstånden mellan mikrofoner och högtalare. Sedan arbetar vi vidare med att skapa kartor med hjälp av sådana

mätningar, och ju mer exakta mätningarna är, desto bättre kommer de slutgiltiga kartorna att vara. Slutligen, om vi har flera sådana kartor – som består av 3D-punktmoln – så visar vi hur dessa kan slås samman till en, mer exakt karta. De lokala kartorna kan vara skapade antingen med hjälp av ljudmätningar eller bilder, som i exemplen ovan.