Seende och tolkning
Artificiellt seende -- robotik

Gösta Granlund

Att kunna se är ett oerhört kraftfullt hjälpmedel för att kunna uppfatta vad som händer i vår omgivning. Det används för att exempelvis navigera i omvärlden, för att hitta föremål, för att undvika faror, för att känna igen människor och kunna identifiera deras sinnestillstånd av välvilja eller fientlighet. Det är därför inte förvånande att en mängd forskning bedrivs i världen för att få fram mekanismer för artificiellt seende, som kan användas åtminstone i enklare situationer, för seende i bl.a. farliga och rutinbetonade tillämpningar.

Att få fram mekanismer för artificiellt seende har dock visat sig svårare än vad man ursprungligen tänkte sig, och vad man skulle ledas att tro, givet datateknikens stora framsteg under åren. Ett problem är att dagens datorer är organiserade för en effektiv bearbetning av symboler arrangerade i strängar. Visuell information däremot ges av spatiala relationer mellan brusiga särdrag, vilka skall testas mot ett stort antal hypoteser eller modeller. Detta är en typ av problem som den konventionella datorns organisation ej är lämpad för och det utgör vad matematikerna kallar ett inversproblem.

Forskarna har därför börjat titta på hur biologiska synsinnen fungerar, för att kunna ta idéer från dessa för att få fram bättre mekanismer för artificiellt seende. Idéerna tas från neurofysiologi, perceptionspsykologi, kognitionsteori, m.fl. områden. Det finns ett antal begrepp som används inom detta område såsom, seende, kognition, perception, artificiell intelligens, m.fl., vilka är delvis överlappande och vars exakta definition beror på vilken subkultur dess användare tillhör. Vi kommer här att använda dessa begrepp relativt synonymt och utan någon detaljerad definition. Man är dock som regel överens om att ***percept*** innebär vissa särdrag i bilden såsom kanter, linjer, objektkonturer, färger, etc. I de flesta fall där de diskuterade mekanismerna är inspirerade från biologiska synsinnen, kommer vi inte att speciellt ange om vi diskuterar ett biologiskt synsinne eller ambitionerna för ett artificiellt.

Ändamålet med ett seende system är att generera ändamålsenliga ***responser*** på vissa percept. Responser kan vara fysiska aktiviteter ut mot systemets omgivning, såsom att flytta på ett föremål. Responser kan också vara en kommunikation till ett annat system att

utföra den fysiska aktiviteten. Responser kan dessutom vara signaler till det egna systemet, att förändra dess bearbetningsmodeller som förberedelse för ett nästa steg.

Man kan särskilja två olika ansatser av mekanismer för artificiellt seende:

1. Föreskrivande av operationer
2. Explorativ inlärning

Den klassiska ansatsen är att försöka föreskriva vad systemet skall göra när vissa kombinationer av percept uppenbarar sig. Detta fungerar för väldefininerade problem med begränsad komplexitet.

Allt eftersom system skall kunna hantera allt mera komplexa situationer framstår dilemmat att programmeraren skall kunna förutse alla situationer som ett system kan råka in i, och vi begränsar systemets kompetens till att klara av ”just vad det har programmerats till att göra”.

Forskarna blir därför allt mera övertygade om att vägen mot kraftfulla seende system oundvikligen går via ***explorativ inlärning***. Det innebär att systemet förses med vissa basfuntioner av percept och mekanismer för rörelse. Sedan får systemet självt utforska sin omgivning, självt lära sig hur saker fungerar. Det är väsentligen hur ett barn lär sig att förstå vad det ser. Det rör sin hand och ser att något händer i dess synfält. Det kommer på att vad det ser, beror på dess egen rörelse av handen. Det kommer så småningom att förknippa eller ***associera*** rörelsen hos handen med rörelsen hos ett objekt i dess synfält. Detta anser man vara en viktig komponent i skapandet av vår rumsuppfattning.

Hur sker sedan sammankopplingen av percept med olika rörelser eller aktioner? Man trodde ursprungligen att objekt i vårt synfält mirakulöst lyckades koppla ihop sig med korresponderande funktionella aktionsmoder till att bygga upp beteenden, styrt från perceptsidan. Situationen är dock att inom ett synfält kan det samtidigt finnas tusentals olika särdrag, delobjekt, objekt, som bombarderar bearbetningsdelen. Att organisera alla dessa komponenter utan vidare och kunna komma fram till hur de hänger ihop, är helt ogörligt. Det utgör ett kombinatoriskt problem med tusentals frihetsgrader.

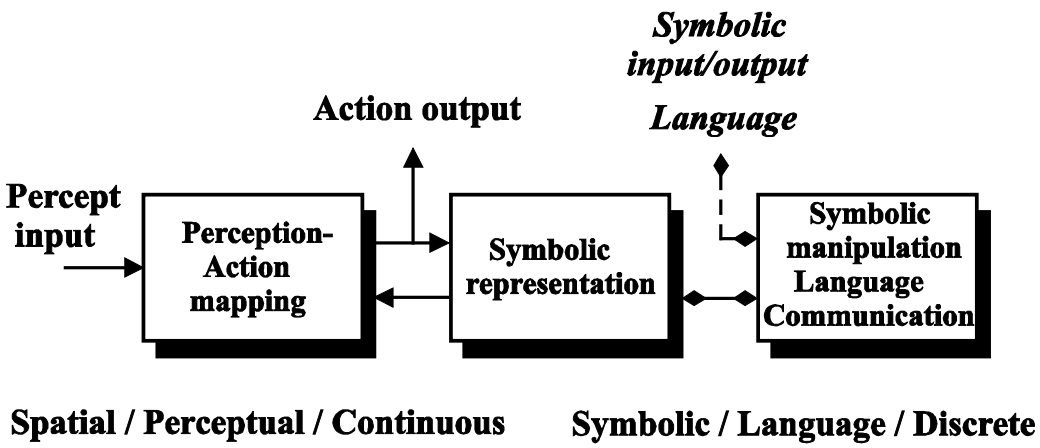
Allt talar för att organisationen sker från aktions- eller responssidan. Det kan karakteriseras som att vid inlärning, ***aktion föregår perception***. Detta ger ett antal fördelar:

1. Antalet frihetsgrader i aktionsrymden är mycket låg. Vi kan exempelvis bara röra oss i tre dimensioner.
2. Problemet förenklas på olika sätt genom att en aktion fokuserar på vissa delar av synfältet. Genom att peta på något som kan vara ett objekt, är det möjligt att ta reda på vilka delar som ingår i objektet, dvs som rör sig mot bakgrunden.

Det visar sig att den viktigaste komponenten för att bygga upp en omvärldsförståelse eller rumsuppfattning i vid bemärkelse, vilket bl.a. innefattar att den innehåller objekt som är fasta eller rörliga, små eller stora, är just rörelseförmågan. Sedan behöver den kopplas samman med en sensormekanism av något slag, som inte nödvändigtvis är syn utan exempelvis känsel. Vi vet att personer som varit blinda från födseln utan problem bygger upp en helt adekvat rums- eller omvärldsuppfattning, även om synen som sensorkanal är överlägsen då den har större kapacitet och möjliggör fler detaljer.

Innebär det föregående att framtidens avanceradesystem för artificiellt seende måste ha rörelseförmåga för att kunna interagera med sin omgivning, och lära sig från denna? I princip ja. Detta skulle i en strikt uttolkning innebära att intelligenta system endast skulle kunna utvecklas i form av robotar. Det verkar dock rimligt att kunna tillhandahålla virtuella omgivningar som ett system under upplärning kan interagera med. Det innebär t.ex. att när systemet ”tänker sig” att det vrider på ett objekt, känns dessa aktionssignaler av från simulatorn för den virtuella omgivningsmodellen, och systemet under inlärning får sig tillhanda en bild med objektet vridet i samma utsträckning. På så sätt kan man tänka sig att steg för steg bygga upp de kompetenser som erfordras för något visst tillämpningsområde.

De mekanismer som beskrivits i föregående avsnitt, hör väsentligen till den spatiala och kontextkänsliga delen av ett seende system. För att ett system för seende skall bli tillräckligt flexibelt fordras även mekanismer för bearbetning av symbolisk information. Vi kan se hur ambitionen att få fram ett tillräckligt flexibelt och läraktigt system leder oss till att steg för steg vara tvungna att inkludera flera mekanismer, såsom rörelseorgan och symbolisk bearbetning, som vi inte trodde krävs ”bara” för ett seende system. Vad vi nu kommer in på är dilemmat att seende inte är bara en avskiljbar process, utan att den är integrerad med hela den uppfattning av omvärlden som den skall betjäna. Hela detta komplex är vad som ofta betecknas kognition.



För att åstadkomma detta kan vi tänka oss ett system med två huvuddelar samt en sammankopplande del, såsom illustreras i vidstående figur. Det vänstra blocket utgör vad som har behandlats i föregående avsnitt. Det arbetar med information i en mycket spatialt kontextkänslig form, och kan sägas hantera ***här-och-nu*** komplexet. Kontextkänsligheten innebär t.ex. att om vi skall trycka på en knapp på en box för att sätta igång en apparat, är det nödvändigt att systemet först klarar av att positionera sig framför boxen, sedan röra handen så att den befinner sig ovanför boxen, sedan positionera fingret över knappen. Först då är kontexten för den egentliga aktionen --- en knapptryckning --- definierad. En knapptryckning i någon annan kontext är naturligtvis meningslös, och ger i vilket fall inte avsett resultat. Vi kan se att ett väsentligt problem i seende eller kognition är just att etablera vad som är korrekt kontext för ett nästa steg, i vad som är blir en stegvis instyrningsprocess. Vi kan också se att denna instyrning måste ha en relativt god noggrannhet, vilket kräver en god koppling mellan rörelseorgan och sensororgan, säg syn. Därav följer den i figuren illustrerade nära kopplingen mellan vad som i figuren betecknas som aktionsutgång och inkommande percept.

Kontextkänsligheten hos den första delen av systemet gör denna mycket tung beräkningsmässigt, och den tenderar att ”sitta fast” i just den kontext där den råkar befinna sig. Det innebär att den inte lätt kan flytta över till en annan kontext annat än att hela systemet befinner sig just över denna. Detta är vad som tidigare beskrivits som dess här-och-nu egenskap. Den kan bara hantera just det lokala området och i nutid.

Detta är funktionellt för vissa situationer, men det är inte tillräckligt om vi vill kunna ***generalisera***, vilket innebär att vi vill att systemets inlärda kompetens skall kunna utnyttjas inte bara just i den punkt där den lärde sig något, utan i andra liknande situationer. I figuren

finns därför längst till höger en del som bearbetar symbolisk information, och typiskt hanterar uppgifter som planering, språk och kommunikation. Symbolisk innebär i vår terminologi information som är fri från detaljerad spatial information och därmed är vad som betecknas relativt ***invariant*** eller oberoende av kontext. Till varje symbol finns dock alltid någon kontext representerad i någon form.

Processen blir nu att den kontextkänsliga informationen i vänstra delen frigörs från sin specifika spatiala kontext till att anta symbolform. Dessa symboler kan nu manipuleras relativt fritt på olika sätt, för att gälla andra positioner, andra tidpunkter oberoende av var systemet råkar befinna sig. Den på så sätt genererade symboliska informationen kan sedan föras från höger till vänster, aktuell kontext kan sättas in och systemets vänstra del kan göra en uppgift i sitt nuvarande läge, med kunskap inlärd i en annan position och tidpunkt.

Mycket talar för att informationen som omvandlas till symbolisk form representerar de aktioner som den första delen genererar. Vi har själva en illusion av att vi observerar den omgivande världen som en förutsättningslös, objektiv realitet genom ett fönster. Det är dock väl känt från kognitionspsykologin att vår medvetna uppfattning av omvärlden ges i termer av de aktioner som vi kan utföra på omvärlden. Dessa aktioner eller aktionsalternativ tolkas sedan och ges en symbolisk form, vilken är vad som utgör vår medvetna uppfattning av omvärlden.

Vi kan något vårdslöst karakterisera den vänstra delen av systemet som en högerhjärna och på motsvarande sätt den högra delen som en vänsterhjärna, mot bakgrund av delarnas specialisering på spatial respektive symbolisk information. Att göra en mera detaljerad jämförelse av dessa koncept, för dock för långt i detta sammanhang.

Tack vare att den symboliska informationen är relativt kontextfri, lämpar den sig att sättas samman i paket och överföras till ett annat system, t.ex. i form av vad vi vardagligen kallar språk. Där tas den om hand och relateras till den information som det mottagande systemet har om omvärlden. Det är viktigt att komma ihåg att informationsinnehållet i språket som sådant är mycket begränsat. Kraftfullheten hos språklig kommunikation kommer av att språkets symboler kan aktivera komplexa strukturer av kunskap om omvärlden som systemet har tillgodogjort sig genom den typ av explorativ inlärning som tidigare beskrivits. Den detaljerade kunskapen om omvärlden måste således redan finnas i systemet, medan den symboliska informationen gör det möjligt att referera till och hantera helt nya kombinationer av dessa redan kända komponenter.

Hur kan då ett system av den här diskuterade typen tillgodogöra sig information? Det verkar finnas tre sätt:

1. Kopiering vid den tidpunkt då systemet genereras
2. Explorativ inlärning, vilket definierar begreppsrymder genom associativ inlärning
3. Passiv observation eller kommunikation, t.ex. med användning av språk

Det är i princip möjligt att kopiera ett existerande system för att generera ett nytt system som är identiskt med det första. Det kan mycket väl finnas praktiska komplikationer som gör det svårt i ett givet fall. Orsaken till att detta är principiellt möjligt är att ingen förståelse av strukturen fordras, utan endast en ”blind” men felfri kopieringsprocedur. Ett exempel på en anordning av denna typ är en vanlig TV-apparat som avbildar punkterna i en bild från en yta till en annan, utan att behöva bekymra sig om vad bilden föreställer. Hos biologiska system har den ”praktiska komplikationen” med kopieringen lösts genom att man varje gång går tillbaka till ritningen, dvs den genetiska informationen om systemet. Medan en kopiering av ett helt system är i princip möjligt, visar det sig vara i princip omöjligt att bara kopiera en del av systemet som innehåller en viss information.

Explorativ inlärning, är vad som har diskuterats i större delen av denna artikel, och denna process generar vad som brukar kallas begreppsrymder, vilka beskriver samband mellan vad ett system gör och vad det observerar som konsekvens av detta. Denna koppling verkar utgöra vad vi vanligen benämner förståelse av ett fenomen på låg nivå.

Efter att ha definierat någon uppsättning av dessa begreppsrymder, kan systemet genom passiv observation ”förstå” objekt eller situationer som använder sig av dessa begreppsrymder, genom interpolation eller extrapolation av dessa. Systemet, eller vi som system, kan förstå en situation som vi inte har erfarit tidigare, så länge som situationen innehåller komponenter som tillräckligt väl liknar komponenter av vilka vi har en explorativ erfarenhet. Detta möjliggör sedan förståelse från passiv observation eller förståelse från kommunikation, t.ex. med användning av språk.

En konklusion av det föregående är att en utveckling av avancerade system för artificiellt seende kräver att dessa har en organisation som möjliggör att de kan lära sig själva från omgivningen. Först då kan tillräckligt av detaljerad information göras tillgänglig för att systemen skall kunna orientera sig med tillräcklig precision i sin kontext. Det är också nödvändigt för att system skall ha en möjlighet att generalisera, dvs kunna klara av något annorlunda och mera komplicerade situationer än exakt vad de har träffat på tidigare.

Föregående presentation är av naturliga skäl ofullständig i ett antal avseenden. Mera information kan återfinnas under:
http://www.cvl.isy.liu.se/Research/

Ett system för artificiellt seende enligt principerna ovan är under utveckling i ett EU-projekt COSPAL, inom det Europeiska forskningsprogrammet IST:
http://www.cospal.org/