Radiosity för teknologer

Stefan Gustavson, ITN (stegu@itn.liu.se) 2005-02-02, uppdaterat 2010-01-29

Radiosity är en matematisk modell av de verkliga mekanismerna bakom verkligt ljusflöde och belysningsintensitet. Man modellerar ljuset i en scen i termer av en energibalans. Modellen är bra och kan ge noggranna simuleringar av verkligheten, även om den har vissa begränsningar. En begränsning är att endast diffusa reflexioner modelleras, men många verkliga scener har nästan enbart diffusa ytor. En annan begränsning är att metoden fortfarande kräver litet för mycket beräkningar för att vara gångbar för rutinmässig daglig användning. Metoden har dock på sistone använts i flera uppmärksammande produktioner, så det är inte längre helt uteslutet att rendera även animerat material med radiosity.

För att närmare förklara både fördelarna och nackdelarna med radiosity för en tekniskt kompetent publik med kunskaper i algebra är det faktiskt enklast att börja med att härleda matematiken bakom metoden. För- och nackdelar ger sig sedan självklart av ekvationerna.

Vi vill modellera diffusa reflexioner i en scen, både direkta reflexioner från ljuskällor och sekundära reflexioner mellan diffusa ytor i scenen. Vi börjar med att dela upp scenen i N små ytelement. Ofta kan man direkt använda de trianglar som använts för modelleringen av scenen, men ibland behöver man dela upp stora trianglar i ytterligare mindre bitar. Ytelementen i en scen som skall renderas med radiosity bör inte vara alltför stora. Detta gäller speciellt i områden nära ljuskällor eller i skuggpartier där man kan förvänta sig en hög kontrast i bilden. Det är också viktigt att man har en sammanhängande sluten yta på sina objekt. Helst skall det inte finnas några objekt eller ytor som skär varandra i scenen, utan allt skall vara korrekt modellerat, hopfogat och placerat. En vanlig anledning till att folk överger radiosity som metod är att det ställer extra höga krav på kvaliteten hos modellerna i en scen. Det duger inte riktigt att det ser bra ut, det skall vara korrekt gjort också.

Ljus från ett ytelement kommer att landa på andra ytelement och påverka dem. Element som ligger nära kommer att påverkas mycket, element som ligger långt ifrån kommer att påverkas mindre, och många element är skymda och kommer inte att påverkas alls. Vi kan generalisera detta och beräkna ett antal så kallade *formfaktorer*. Formfaktorn F_{ij} anger hur stor andel av ljuset som sänds ut från element j som kommer att belysa element i, eller något förenklat: hur mycket av element i som är synligt sett från ytan av element j.

Ljuskällor modelleras som självlysande element i scenen, och dessa ges en egenemission ε_i. För varje element i scenen kan då det direkta ljuset I_o endast från själva ljuskällorna tecknas:

$$I_{0,i} = \varepsilon_i$$

Detta ger förstås ingen särskilt intressant eller realistisk bild, eftersom bara ljuskällorna är synliga och allt annat blir svart. Om vi vill modellera både direkt emission och den första diffusa reflexionen av ljus från ett element behöver vi ta hänsyn till följande:

- egenemissionen ε_i från elementet självt
- formfaktorerna F_{ij} som anger hur mycket av elementet i som är synligt från vart och ett av de andra elementen j
- emissionen från alla andra element ε_j
- reflexionskoefficienten R_i för elementet (det vill säga ytans diffusa färg).

Det direkta ljuset samt den första reflexionen kan då modelleras enligt följande:

$$I_{1,i} = \varepsilon_i + R_i \sum_{j=0}^{N} \varepsilon_j F_{ij}$$

Detta är ungefär vad en lokal ljusmodell för diffus reflexion klarar av, så det ger inte heller särskilt realistiska bilder. Skuggor som inte är alltför skarpa hanteras korrekt genom att beräkna formfaktorerna noggrant, men i övrigt blir ljuset ganska platt och tråkigt. De diffusa interreflexionerna mellan objekt saknas, och det är de som ger det mjuka tilltalande utseendet för en verklig diffust belyst scen.

Det är dock inga problem att utöka modellen till fler nivåer av interreflexion. Om vi vill ta hänsyn till även den andra ordningens reflexion, alltså reflexionen från ljuskällor via en diffus yta till en andra diffus yta, så kan vi direkt modellera en reflexion till med samma slags ekvation som vi redan har använt:

$$I_{2,i} = \varepsilon_i + R_i \sum_{j=0}^{N} I_{1,j} F_{ij}$$

Notera att eftersom ε_i ingår i $I_{1,i}$ så kommer summan här att inkludera både första och andra ordningens reflexioner.

Vi kan generalisera detta till godtyckligt många nivåer av interreflexioner. Om vi vill ta hänsyn till upp till n stegs upprepad reflexion mellan diffusa ytor så kan vi teckna:

$$I_{n,i} = \varepsilon_i + R_i \sum_{j=0}^{N} I_{n-1,j} F_{ij}$$

Det som beskrivits ovan är faktiskt en fullt gångbar och ofta använd lösningsmetod för radiosity-beräkningar. Den kallas för successiv approximation eller ”progressive refinement” och medger att man avbryter renderingsprocessen när man inte längre ser någon skillnad mellan en nivå och nästa, alltså när det är så litet ljus som fortfarande studsar omkring att det inte representerar någon väsentlig del av den totala energin i scenen.

Det slags iteration som representeras av ekvationen exakt som den står brukar kallas för ”gathering”, eftersom man för varje element i samlar ihop inverkan från varje annat element j. Man kan även göra beräkningen i andra ordningen, så att man för varje element j sprider dess reflekterade ljus till alla andra element i som påverkas, och därefter fortsätter med nästa element. Man har då möjligheten att börja med att räkna på de element som sänder ut mest ljus och eventuellt strunta i dem som inte sänder ut något i nämnvärd grad. Detta brukar kallas ”shooting” och är också en vanlig iterativ lösningsmetod för radiosity.

Man kan dock teckna en exakt lösning till ekvationen i stället för att fuska med iterativa metoder. Vi börjar med att teckna ekvationen ovan som en matrisekvation, vilket gör notationen alldeles oerhört mycket enklare. Vi samlar ljusintensiteterna och emissionerna för de N elementen i vektorer $\overline{I_n}$ och $\bar{\varepsilon}$ med N element vardera och låter formfaktorerna bilda en $N \times N$-matris F där F_{ij} är elementet på rad i, kolumn j. Reflexionskoefficienterna lägger vi längs diagonalen i en $N \times N$ diagonalmatris R, där R_i är elementet på rad i, kolumn i, och övriga element i R är noll. Ekvationen för samtliga elements intensitet efter n reflexioner kan då tecknas:

$$\overline{I_n} = \bar{\varepsilon} + RF\overline{I_{n-1}}$$

Nu observerar vi att det vi är intresserade av är det *stationära tillståndet* i systemet, alltså gränsvärdet när det inte är någon skillnad mellan ett steg och nästa av simuleringen. På grund av fysikaliska villkor så har en verklig scen alltid en energibalans. Denna balans uppnås i verkligheten på bråkdelar av en mikrosekund när man tänder, släcker eller flyttar ljuskällor, och eftersom ytor enligt termodynamikens första lag inte kan reflektera mer energi än de tar emot så har vi en garanterad konvergens i systemet. Om formfaktorerna F_{ij} är korrekt beräknade och vi har fysikaliskt rimliga reflexionskoefficienter $0 < R_i < 1$ så har vi också en garanterad konvergens i vår iterativa lösning, vilket innebär att vi har ett gränsvärde:

$$\overline{I_n} \to \overline{I_{n-1}} \text{ då } n \to \infty$$

Om vi nu helt fräckt antar att vi nått gränsvärdet och kan sätta $\overline{I_n} = \overline{I_{n-1}} = \bar{I}$ så får vi:

$$\bar{I} = \bar{\varepsilon} + RF\bar{I}$$

Om vi stuvar om litet mellan höger- och vänsterledet och inför en enhetsmatris E får vi:

$$(E - RF)\bar{I} = \bar{\varepsilon}$$

och den sökta lösningen kan direkt tecknas genom att invertera matrisen $(E - RF)$:

$$\bar{I} = (E - RF)^{-1}\bar{\varepsilon}$$

Vi har här en sluten lösning som utifrån emissionen $\bar{\varepsilon}$ hos de lysande ytorna direkt beräknar den reflekterade ljusintensiteten från varje element i scenen, *utan* att vi angett någon speciell kamerariktning. Vi har löst en generell ekvation för energibalansen hos ljuset i hela scenen. Om vissa eller rentav alla ljuskällor ändrar sin intensitet kan vi enkelt räkna ut inverkan av detta utan att lösa ekvationen från början - det räcker med att multiplicera den nya ljusvektorn $\bar{\varepsilon}$ med den redan beräknade matrisen $(E - RF)^{-1}$.

Detta visar på generaliteten hos radiosity-metoden: genom att beräkna en matris av formfaktorer och invertera den kan vi förberäkna belysningen för samtliga ytor i scenen, och kan rendera en korrekt ljussatt bild från vilken vinkel som helst utan räkna på något mer än vilka ytor som är synliga och var de hamnar i bilden, alltså inte mer jobb än för en mycket enkel rendering med en konstant färg hos varje yta. Med en förberäknad radiosity-lösning kan alltså scenen renderas mycket snabbt, även i realtid, och med litet planering kan man även hantera att vissa ljuskällor tänds och släcks eller ändrar intensitet. Sådana förberäkningspass görs numera ofta i datorspel, och man lagrar då ljusinformationen för scenen som separata texturbilder, så kallade ”light maps”.

Slutekvationen visar samtidigt på bristerna hos metoden. För att få fram en bild måste vi beräkna ett stort antal formfaktorer F som beror på scenens geometri, och detta kan i sig vara ett mycket omfattande arbete. Dessutom måste vi invertera en matris som för en scen med N element innehåller $N \times N$ koefficienter. I en scen med hundratusentals element, vilket inte på något sätt är extremt mycket, kommer matrisen snabbt att bli ohanterligt stor och i stort sett omöjlig att invertera utan en rent orimlig beräkningskraft. Det är trevligt att veta att det *finns* en exakt lösning, men den är sällan möjlig att ta fram. Den iterativa, approximativa lösningen har flera fördelar. Dels behöver man inte invertera någon jättematris, men dels behöver man inte heller lagra alla formfaktorer. Formfaktorer F_{ij} som är noll eller mycket små kan man strunta i, vilket gör mängden koefficienter i F som verkligen behöver lagras mer rimlig, även för stora scener. Att lösa ekvationen iterativt ger också möjligheter att rimligt snabbt få en approximativ lösning som sedan förfinas i varje steg, i stället för att vänta en halv evighet på att få en exakt lösning och dessförinnan inte ha någon lösning alls.

En grundläggande brist hos radiosity som renderingsmetod är att den inte lämpar sig för scener där det sker väsentliga förändringar av objektens position. Om något objekt flyttar på sig så påverkas formfaktorerna inte bara för det objekt som rör sig och dess direkta inverkan på andra objekt. Det objekt som flyttar på sig påverkar ett stort antal formfaktorer på ett ganska oförutsägbart sätt, eftersom formfaktorerna även tar hänsyn till vad som skyms av det rörliga objektet. Radiosity är därför mest lämpad för statiska eller nästan statiska scener.

Radiosity har dock visat sig kunna ge mycket realistiska bilder, och det är definitivt mer än en akademisk övning i ekvationslösning. Det används och gör nytta i en rad sammanhang.