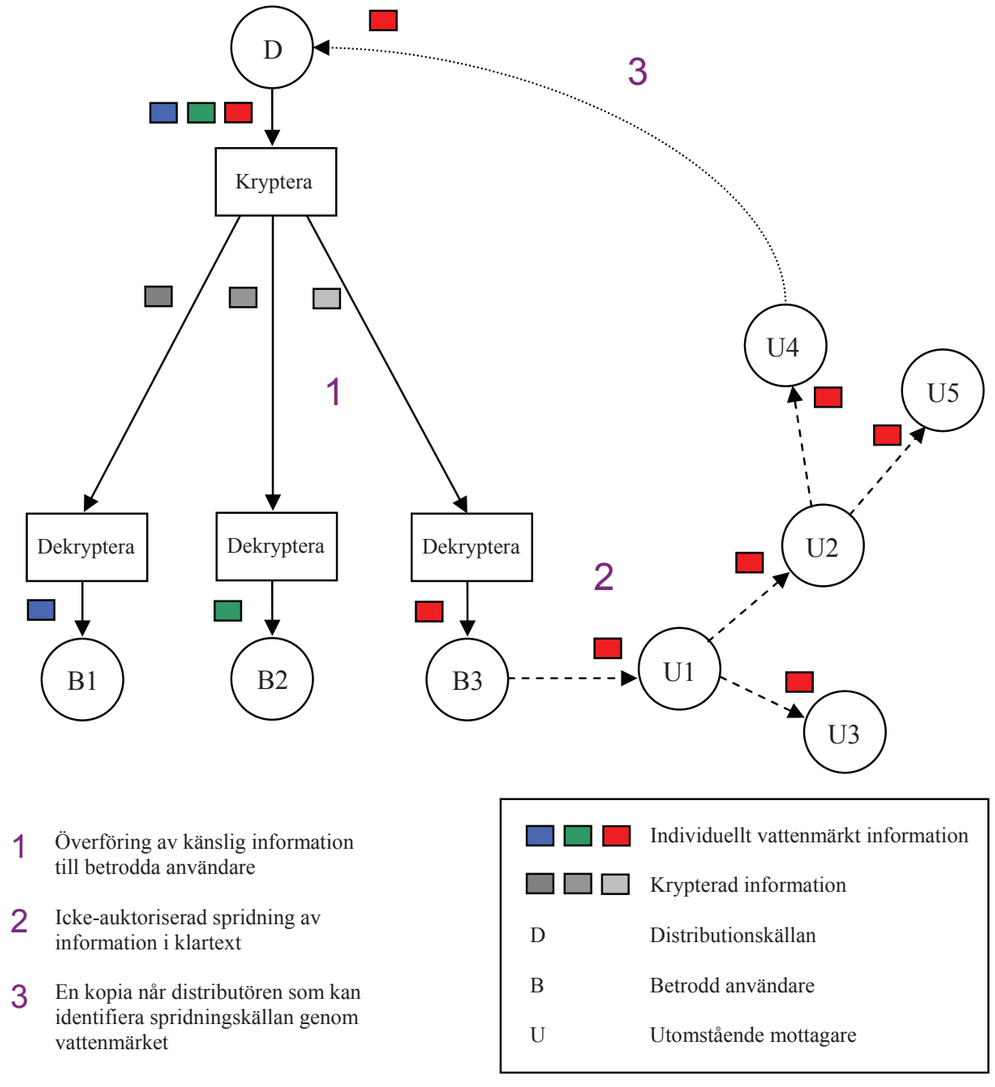
Vattenmärkning av sensordata

Slutrapport

K-G STENBORG, JONAS ALLVAR, GUSTAV HAAPALAHTI,
MAGNUS HERBERTHSON



FOI
Totalförsvarets forskningsinstitut
Informationssystem
Box 1165
581 11 Linköping

Tel: 013-37 80 00
Fax: 013-37 81 00

www.foi.se

FOI-R--3138--SE Underlagsrapport **Informationssystem**
ISSN 1650-1942 December 2010

K-G Stenborg, Jonas Allvar, Gustav Haapalahti,
Magnus Herberthson

Vattenmärkning av sensordata

Slutrapport

Titel	Vattenmärkning av sensordata - Slutrapport
Title	Watermarking of sensor data – Final Report

Rapportnr/Report no	FOI-R--3138--SE
Rapporttyp Report Type	Underlagsrapport
Sidor/Pages	37 p
Månad/Month	December
Utgivningsår/Year	2010
ISSN	ISSN 1650-1942
Kund/Customer	FM
Projektnr/Project no	E3093
Godkänd av/Approved by	Lars Bohman

FOI, Totalförsvarets Forskningsinstitut	FOI, Swedish Defence Research Agency
Avdelningen för Informationssystem	Information Systems
Box 1165	Box 1165
581 11 Linköping	SE-581 11 Linköping

Sammanfattning

Sensordata som innehåller ljud, bild och video är viktiga, och kommer i framtiden vara ännu mer betydelsefulla, inom försvarsmakten. Därför är upprätthållandet av sekretess och integriteten av denna data väsentlig. Vattenmärkning (Watermarking) är en metod som kan användas för att bädda in information i sensordata. Efter inbäddningen skall det inte finnas någon perceptuell skillnad mellan originaldata och den sensordata som fått ett digitalt vattenmärke. Detta vattenmärke kan vid ett senare tillfälle utvinnas ur sensordatat och användas för olika säkerhetstillämpningar.

Inom Strategiska forskningskärnan *Vattenmärkning av sensordata* har tre olika former av digitala vattenmärken undersökts:

- **Identifikationsmärkning** som används för att lagra information om sensordatats ursprung och identitet. På så vis följer denna information alltid med sensordata även om den konverteras till andra format.
- **Individuella vattenmärken** som kan användas för att spåra källan av otillåtet spridd information. I dagsläget är just spridning av hemligt material ett ökande problem. Sådan information kan exempelvis hamna på hemsidor som Wikileaks och Openleaks.
- **Förändringsindikation** som är en form av vattenmärkning som används för att undersöka om sensordata är original eller om det har blivit ändrat. Det går även se inom vilka områden som exempelvis en bild har blivit förändrad. Dessutom går det bygga in metoder för att delvis kunna återskapa de skadade områdena.

Rapporten innehåller beskrivningar om de olika formerna av vattenmärkning. Den redovisar relationen mellan informationsmängd i vattenmärket, robusthet och den distorsion som vattenmärket inför i sensordata. Lösningar redogörs för hur individuellt vattenmärkt sensordata på ett effektivt sätt skall överföras och lagras. Den demonstrationsmetod för förändringsindikation som har implementerats presenteras.

Inbäddningen av vattenmärken är anpassad för att distorsionen som uppstår inte skall vara tydliga för det mänskliga synsinnet. Däremot behöver det inte betyda att de är lämpade för automatiska signalbehandlingsmetoder. Påverkan på en utvald måldetektionsalgoritm har undersökts. Det visar sig att så länge som ett inte orimligt starkt vattenmärke används, ger vattenmärkningen ingen större negativ påverkan på detektorn.

Nyckelord: vattenmärke, vattenmärkning, vattenstämpel, digital signatur, watermark, watermarking, fingerprinting, autentisering

Summary

Sensor data containing audio, image and video are important, and will in future be even more significant, within the Swedish Armed Forces. Therefore, maintenance of confidentiality and integrity of this data material is essential. Watermarking is a method that can be used to embed information in the sensor data. After the embedding, there should be no perceptual difference between the original data and the watermarked sensor data. The watermark can at a later stage be extracted from the sensor data and used for security applications.

Within the project "Watermarking of sensor data" three different types of digital watermarks has been considered:

- **Identification** that is used to store information about the sensor data origin or identity. This information will always follow the sensor data, even if it is converted to other formats.
- **Individual watermarks** can be used to trace the source of illicit spread information. Today spreading of secret material is an increasing problem. Such information can for example be spread to websites such as Wikileaks and Openleaks.
- **Content authentication** that is a form of fragile watermarking used to verify the integrity of the sensor data. This can also be used for locating which areas for example an image has been altered. Additionally, you can build in methods to partially restore the damaged areas.

This report contains descriptions of the different forms of watermarking. It presents the relationship between the amount of information in the watermark, robustness and distortion that the watermark causes to the sensor data. How individually watermarked sensor data effectively be transferred and stored has been investigated. The demonstration method for content authentication that has been implemented is described.

Embedding watermarks are adapted so that the distortion that occurs in the sensor data will be obscure to the human visual sense. Automatic signal processing can be affected by other parameters compared to humans. How a selected detection algorithm is affected if the sensor data has a watermark embedded has been investigated. It turns out that as long as not an unreasonably strong watermark is used, no significant negative impact on the detector from the watermark can be noticed.

Keywords: watermark, watermarking, fingerprinting, digital signature, authentication

Innehållsförteckning

1 Inledning

Detta kapitel innehåller en kortfattad beskrivning av digitala vattenmärken, samt en lista på de publikationer som gjorts inom projektet *Vattenmärkning av sensordata*.

1.1 Bakgrund till vattenmärkning

Vattenmärkning (Watermarking) är ett sätt att bädda in information i olika typer av sensordata som bilder, ljud och video. Namnet vattenmärke[1] kommer från den typ av märke som sedan flera hundra år används för att märka viktiga papper såsom sedlar och arkiveringspapper (Hartung, 1999).

Digitala vattenmärken kan användas för att märka olika former av digitala data såsom stillbilder, video och ljud. I denna rapport kommer nästan uteslutande stillbilder användas för att de dels är enkla att arbeta med, men också för att det är väl anpassade för presentation i pappersformatet, till skillnad från ljud och video. Det som gäller för stillbilder kan dock utan större modifiering även användas för video.

Det finns många typer av digitala vattenmärkningsmetoder och många olika användningsområden. I denna rapport har vi tittat närmare på tre olika användnings-områden av vattenmärken:

- **Identifikationsmärkning** är robusta vattenmärken som används för att lagra information, exempelvis om vilken sensor som skapat data eller information om vem som är ägaren av en bild. Mer om detta i kapitel 2.
- **Individuella vattenmärken** är unika för varje användares kopia av sensordata. Om exempelvis en bild har spridit sig vidare olovligt går det att genom att undersöka dess individuella vattenmärke spåra spridningskällan. Mer om detta i kapitel 3.
- **Förändringsindikation** kan uppnås med hjälp utav bräckliga vattenmärken som går sönder om bilden manipuleras. I ett senare skede går det att kontrollera om vattenmärket i sensordata är helt. Om det är skadat vet man att sensordatan inte längre är original. Mer om detta i kapitel 4.

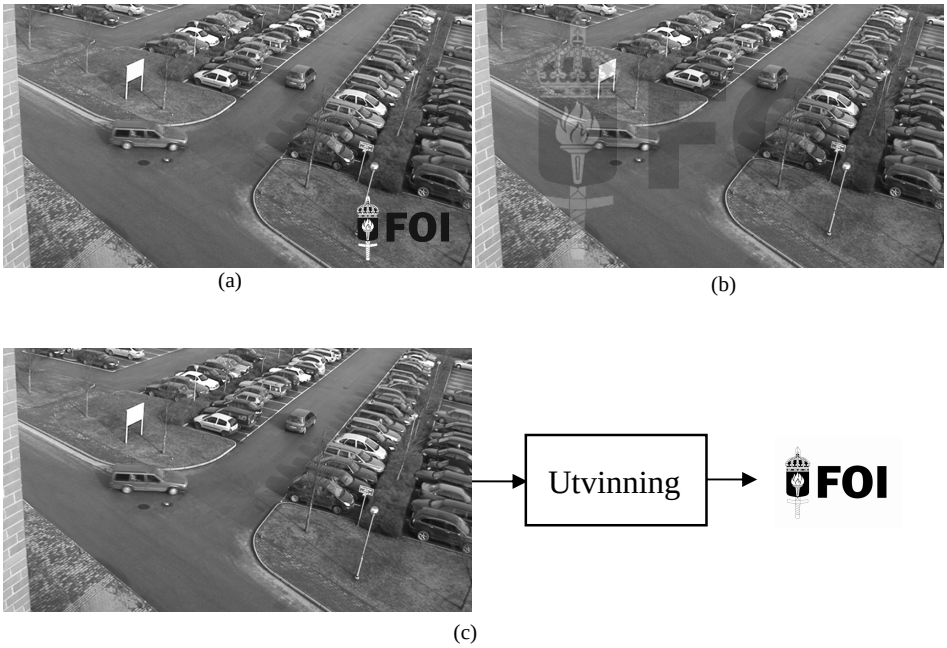
Vi har i första hand koncentrerat oss på visuella sensordata[2] som bilder och video. Vattenmärken kan delas in i två klasser, synliga och osynliga.

- **Synliga** vattenmärken är sådana som tydligt kan ses av alla användare. Vanliga typer av synliga vattenmärken är de copyrighttexter som kan finnas i nederkanten på fotografier eller de tv-kanalslogotyper som vanligtvis ligger i övre högra hörnet av tv-bilden. Ett synlig vattenmärke kan ses i figur 1 (a) och (b).
- **Osynliga** vattenmärken är sådana som inte är synliga för blotta ögat utan ligger inbäddat som till exempel ett brus i bilden. Det är denna typ av perceptuellt osynliga vattenmärken som vi är intresserade av. Ett osynligt vattenmärke illustreras i figur 1 (c).

Denna rapport behandlar osynliga märken, men även synliga märken har många användningsområden.

[1] Vi har valt att använda namnet ”vattenmärke” även om ”vattenstämpel” är ett namn som på svenska är vanligare för den typ av märkning som görs i papper.

2 Med visuella sensordata tänker vi oss här data som representeras visuellt, dvs. även t.ex. IR-data som ligger på för ögat inte uppfattbara våglängder betraktas som visuella eftersom det går att titta på dess sensordatarepresentation.



Figur 1: (a) och (b) Synliga vattenmärken innehållande FOIs logotyp. Denna typ av synlig vattenmärkning är ganska enkel att ta bort. (c) Osynligt vattenmärke som yttrar sig som ett svagt brus i värdbilden, som bara kan noteras när skillnaden mellan originalbild och vattenmärkt bild beräknas. För ett utgående från bilden få tillgång tillvattenmärket krävs en utvinnings-algoritm.

Kryptering och vattenmärkning bör oftast användas tillsammans eftersom de har olika svagheter och styrkor vilket gör att de kompletterar varandra. Kryptering hindrar obehöriga från att till exempel kunna se en bild men det fungerar bara så länge som bilden inte har blivit dekrypterad. Vattenmärkning däremot kan användas för att knyta en otillåtet spridd kopia av en bild till dess källa, eller användas för att se var och när bilden är tagen och vem som äger den, eller för att verifiera att en bild inte har blivit manipulerad. Vilken typ av kryptering och vattenmärkning som skall användas är beroende på vilken typ av säkerhet som vill uppnås:

- **Sekretess**, som betyder att bara rätt personer skall ha tillgång till data. För det används olika intrångsskydd och kryptering.
- **Integritet**, vilket syftar på att data skall vara pålitlig. Detta behandlas alltså av förändringsindikation i kapitel 4.
- **Tillgänglighet**, betyder att informationen skall finnas tillgänglig för rätt person när den behövs. Detta gäller ju även för det till sensordatat tillhörande metadatat, som identifikationsmärkning i kapitel 2 tar upp.
- **Spårbarhet**, betyder att om säkerhetsproblem uppstår skall det i efterhand gå verifiera vad som skett. Detta är en viktig del för individuella vattenmärken i kapitel 3.

En viktig egenskap hos vattenmärkningsmetoder är dess robusthet. Om ett vattenmärke är robust betyder det att vattenmärket fortfarande kan utläsas trots att sensordata har blivit förändrat. Ett bräckligt vattenmärke däremot skadas om bilden som märket är inbäddad i blir skadat. Självklart finns det olika nivå på robusthet. Ett vanligt mål är att vattenmärkningsmetoden skall ge märken som går att tyda så länge som det finns någon användbar information kvar i sensordata. Är sensordata så skadat att det inte längre kan

användas för sina syften är det inte heller lika viktigt att vattenmärken som är inbäddade i datan kan utläsas. I vissa fall är det önskvärt att inte ha robusta utan bräckliga eller halvbräckliga vattenmärken, se kapitel 4.

1.2 Publikationer

Under de tre år som projektet *Vattenmärkning av sensordata* pågått har arbetet kontinuerligt beskrivits i följande publikationer:

- K-G Stenborg, Magnus Herberthson och Gustav Haapalahti *Identifiering och förändringsindikering av sensordata via vattenmärken - en litteraturstudie*, FOI-R--2699--SE, December 2008.
- K-G Stenborg, *Lägesrapport Vattenmärkning av sensordata*, FOI Memo 2695, December 2008.
- K-G Stenborg. *Scalable Distribution of Watermarked Media*, Handbook of Research on Secure Multimedia Distribution, Chapter 18, 335-351, Information Science Reference, IGI Global, Mars 2009.
- K-G Stenborg. *Individually Watermarked Information Distributed Scalable by Modified Transforms*, FOI Memo 2832, Juni 2009.
- K-G Stenborg. *Individually Watermarked Information Distributed Scalable by Modified Transforms*, RTO-MP-IST-087: Information Management and Exploitation, Paper 16, Stockholm, Oktober 2009.
- K-G Stenborg och Gustav Haapalahti. *Vattenmärkning av sensordata - Årsrapport 2009*, FOI-R--2943--SE, December 2009.
- K-G Stenborg och Robert Forchheimer. *Individual Watermarking over P2P Networks*, Technology and Methodology for Security and Crisis Management TAMSEC'10, Oktober 2010.
- K-G Stenborg, Jonas Allvar, Gustav Haapalahti och Magnus Herberthson. *Vattenmärkning av sensordata - Slutrapport*, FOI-R--3138--SE, December 2010.
- K-G Stenborg, Magnus Herberthson och Robert Forchheimer, *Distribution of Individually Watermarked Content in Peer-to-Peer Networks*, in 4th IFIP International Conference on New Technologies, Mobility and Security NTMS'11, Paris, Feb 2011. Kommande publikation.

Denna rapport kommer att ta upp olika delar av det som gjorts under projektets gång. För mer utförlig information av vissa delar hänvisas till tidigare publikationer.

2 Identifikationsmärkning

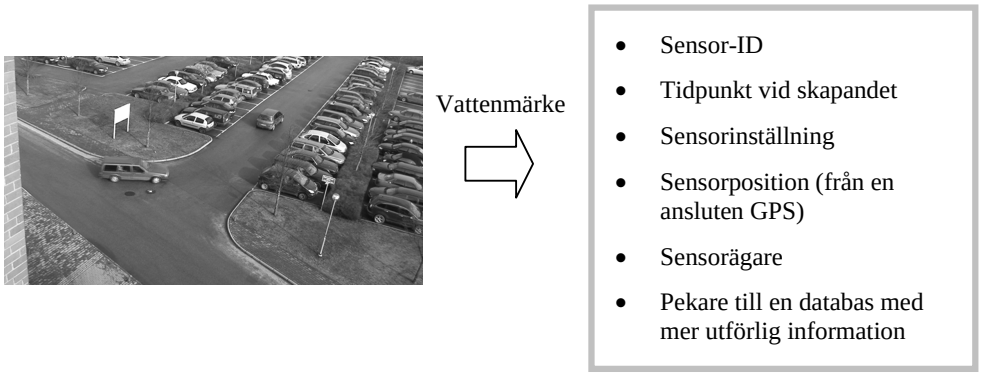
Detta kapitel handlar om robusta vattenmärken som innehåller information om den sensordata märket är inbäddad i. Även förhållandet mellan vattenmärkets robusthet, informationskapacitet och distorsion på sensordata tas upp.

2.1 Exempel på problemställning

Antag att inspelad sensordata, över ett strategiskt viktigt område, finns lagrad på en hårddisk. Problemet är att denna sensordata har genomgått en eller flera formatförändringar vilket har lett till att den headerdata som tidigare fanns lagrad nu har försvunnit. Det medför att det är okänt när data är insamlat, vem som gjort det och exakt vilken utrustning som användes. Därmed har denna sensordata förlorat sin användning.

Ett annat scenario är att synkroniseringen mellan metadata och bilderna i en videoström har gått förlorad. Att i efterhand synkronisera data (exempelvis GPS-data) med videoströmmens tidkod är kostsam.

En metod att undvika att metadata försvinner om headerdata skadas är att bädda in den som vattenmärken, figur 2. För en stillbild betyder det att själva informationen finns lagrad som ett svagt ”brus” på själva pixelvärdena. Liknande inbäddning kan göras för video och där ligger då vattenmärkesinformationen automatisk synkroniserad med bilden.



Figur 2: En bild innehåller ett vattenmärke och detta märke kan innehålla olika typ av information.

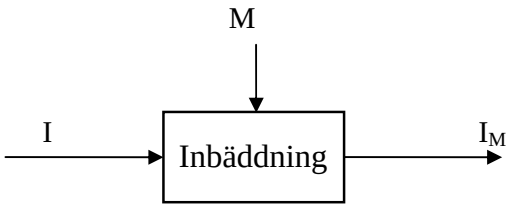
2.2 Robusta digitala vattenmärken

Ofta medföljer det olika former av sidoinformation med insamlade sensordata. För enkelhets skull kommer vi mest att tala om stillbilder i denna rapport, men vattenmärken kan även bäddas in i video, ljud och andra former av sensordata. Den till bilden tillhörande informationen kan exempelvis ange sensorinställningar och tidpunkt för när bilden togs. Denna metainformation lagras ofta i en header till bilden eller i en separat informationsfil. Sådan metadata kan lätt försvinna om exempelvis bilden kodas om till ett annat format. Dessutom är den känslig för yttre påverkan av personer som vill skada och manipulera data. Detta medför att det i ett senare skede kan det vara svårt att avgöra var, när och hur denna bild skapades.

Ett sätt att försäkra sig om att information alltid medföljer en bild är att lagra den som ett inbäddat vattenmärke. Detta vattenmärke måste då vara robust så att det överlever omkodningar och andra händelser som bilden kan råka ut för. I vissa scenarion kan det finnas risk för medvetna attacker där någon medvetet försöker tvätta bort vattenmärket

från en bild. Därför bör vattenmärkningsmetoden vara robust mot olika bildbehandlingsattacker som brus, beskärning, kvantisering, rotation och filtrering.

Det finns många olika algoritmer för att bädda in och utvinna digitala vattenmärken. I (Stenborg, dec 2008) redovisade vi en litteraturstudie på området. I denna rapport ger vi bara en enkel generell beskrivning nedan.



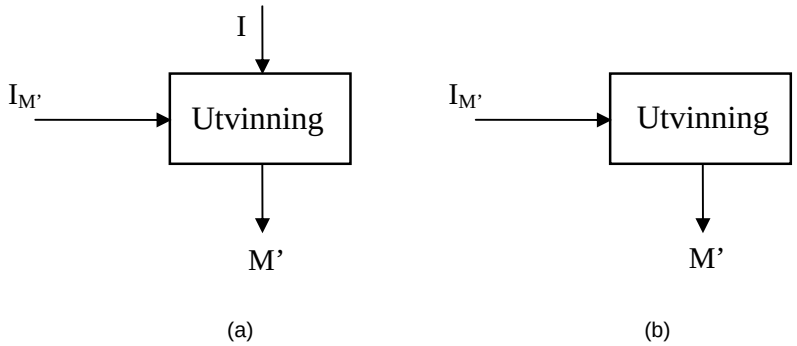
Figur 3: Inbäddning av vattenmärket *M* i bilden *I* vilket resulterar i den märkta bilden I_M.

2.2.1 Inbäddning av robusta vattenmärken

Antag att stillbilden *I* skall vattenmärkas med märket *M*. I så fall skall vi på något sätt ändra lite på pixlarna i *I* så att vårt märke *M* blir inbäddat utan att man med det mänskliga synsinnet ser någon skillnad. Vi kallar bilden med inbäddat vattenmärke för I_M. Utgående från I_M skall vi sedan kunna utvinna märket *M* eller kontrollera om bilden är märkt med märket *M*.

Inbäddningen av märket i bilden kan ske på flera olika sätt. Det går till exempel att bädda in vattenmärket direkt i bilden (spatialdomän) eller efter någon transform, till exempel i krusningsdomänen (även känd som Wavelet eller vågelement) eller i någon frekvensdomän.

Figur 3 ger en generell beskrivning av hur bilden *I* får vattenmärket *M* inbäddat så I_M skapas. Denna och övriga figurer i rapporten är förenklade. Egentligen använder de flesta praktiska metoder nämligen någon form av nycklar vid inbäddning och utvinning av vattenmärkena.



Figur 4: (a) Icke-blind vattenmärkningsutvinning. För att utvinna vattenmärket *M'* ur bilden $I_{M'}$ krävs originalbilden *I*. (b) Blind vattenmärkningsutvinning. Originalbilden behövs inte för att utvinna vattenmärket utan det räcker med den vattenmärkta bilden $I_{M'}$.

2.2.2 Utvinning av robusta vattenmärken

Om vi får en vattenmärkt bild $I_{M'}$ måste vi på något vis få fram vilket märke M' som finns inbäddat. För att kunna utvinna märket kan olika typer av information behövas. Vi kan därför dela in vattenmärkningsmetoderna och deras utvinning i följande klasser:

- **Icke-blind vattenmärkning** (Nonblind Watermarking) – För att kunna utvinna vattenmärket ur en bild $I_{M'}$ krävs originalbilden I. Se figur 4 (a).

- **Blind vattenmärkning** (Blind Watermarking) – Originalbilden behövs inte för att kunna utvinna vattenmärket ur bilden $I_{M'}$. Se figur 4 (b).

Anledningen till att vi kallar den bild som skall undersökas för $I_{M'}$ är att den kanske inte längre är exakt likadan som den inbäddade bilden I_M. Efter inbäddningen kan bilden ha exponerats för skador eller medvetna attacker vilket också orsakar att det utvunna märket M' kanske inte längre är exakt som den inbäddade märket M.

För de flesta tillämpningar är blind vattenmärkning att föredra eftersom originalbilden då inte behövs vid kontroll av vattenmärket. Förutom blind och icke-blind vattenmärkning finns följande klasser för utvinningen:

- **Privat vattenmärkning** (Private Watermarking) – Endast auktoriserade parter kan utvinna vattenmärket ur I_M genom att vattenmärkningsmetoden är hemlig eller genom att en hemlig nyckel krävs för att genomföra utvinningen.

- **Öppen vattenmärkning** (Public Watermarking) – Alla kan utvinna vattenmärket då utvinningsmetoden är känd och ingen hemlig nyckel krävs.

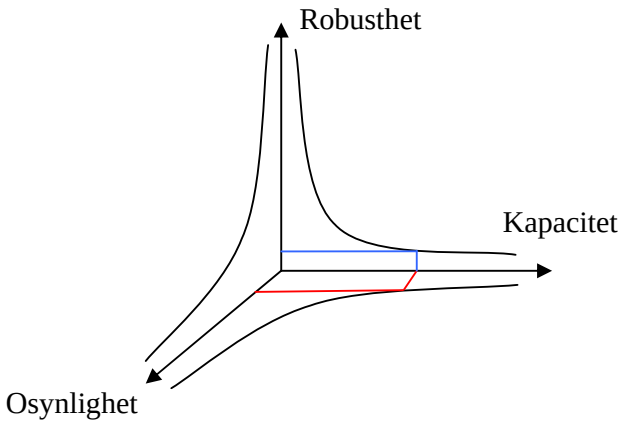
Privat vattenmärkning är ofta att föredra eftersom den oftast anses vara säkrare. Fast vid vissa tillämpningar är öppen vattenmärkning önskvärd, t.ex. om vattenmärket innehåller uppgifter om upphovsrättsinnehavare och man vill att alla skall kunna se vem som är bildens ägare.

2.2.3 Robusthet, kapacitet och distorsion

Mängden information som kan bäddas in som ett vattenmärke i en bild är beroende av bildens storlek (antal pixlar) och vilken robusthet som önskas uppnås (felsannolikhet). Dessutom är det självklart beroende på vilken vattenmärkningsalgoritm som används. Inbäddningen av vattenmärket kommer dessutom påverkas kvalitén på bilden. Ju mer information och högre robusthet på vattenmärket, desto mer distorsion införs det i bilden. Detta illustreras i figur 5 där en hög grad av informationsmängdskapacitet medför låg robusthet och liten osynlighet (= kraftig distorsion). Figur 23 (d) visar ett extremfall för hur ett robust vattenmärke med hög kapacitet kan införa mycket distorsion i värdbilden. Mer utförlig information om robusta vattenmärken som innehåller metadata finns i (Stenborg, 2008) och (Cox, 2008).

Generellt gäller att ju mer sensordata det finns desto mer information kan vattenmärket innehålla. Därmed kan exempelvis hyperspektrala sensordata innehålla mer inbäddad information än en enkel stillbild. En färgbild kan innehålla lite större vattenmärken än en svartvit IR-bild. Samtidigt måste man räkna med hur mycket sensordata som skall behövas för att utvinna vattenmärket. Behövs det en timmes videomaterial eller bara 10 sekunder med video för att utvinna vattenmärket? För hyperspektrala data kan varje band innehålla hela vattenmärket var för sig. I nästa kapitel tar vi upp individuella vattenmärken som används för att spåra icke-auktoriserad spridning. Då är det ofta önskvärt att bara en bild ur en videoström skall räcka för att utvinna vattenmärket. Detta kommer sig av att det är

möjligt att det enda av den spridda videofilen som görs tillgänglig är en enstaka bild på en viktig händelse.



Figur 5: Förhållandet mellan den mängd information ett vattenmärke kan innehålla (kapacitet), hur väl vattenmärket bevaras om värdbilden påverkas (robusthet) samt vilken perceptuell påverkan som vattenmärket ger på värdbilden (osynlighet). Om kapaciteten ligger på nivå där den blå och röda linjen börjar blir robusheten låg (blå linje) och osynligheten blir också låg (röd linje), dvs vattenmärket inför mycket distorsion.

2.3 Påverkan på signalbehandlingsmetod

De flesta nu existerande vattenmärkningsmetoder bygger på att den distorsion som vattenmärket orsakar på värdbilden inte skall uppfattas av det mänskliga synsinnet. Det betyder att det mänskliga synsinnet traditionellt har styrt vilken form av distorsion som är acceptabel. När nu fler och fler automatiska bildbehandlingsmetoder börjar användas på sensordata är det inte länge säkert att denna form av distorsion är den bäst lämpade. En detektionsalgoritm kanske reagerar på andra egenskaper i bilden, än det mänskliga synsinnet.

Vi har därför testat en vattenmärkningsmetod genom att utföra detektioner på bilder innehållandes vattenmärken. Metoden är densamma som presenteras i kapitel 4.

För att testa hur vattenmärkning påverkar detektion så testas en detektionsmetod som bygger på boosting (Näsström, 2009). Boostingmetoden kräver exempelbilder, positiva träningsdata, av de objekt man vill hitta, för att tränas till att detektera de objekten. Det kvävs också en stor mängd bilder av sådant man inte vill hitta, negativa träningsdata. I detta fall är det bilar av en viss typ som ska detekteras. Bilderna är syntetiskt genererade med SceneServer (Bennet, 2003) och vyn är från ett UAV perspektiv. Positiverna genereras inom ett horisontellt vyintervall mellan 25 och 65 grader och en vertikal infallsvinkel mellan 0 och 20 grader. Varje positiv bild (positiv patch) är 30x30 pixlar stor. Bilderna som används för träning är till antalet: 5175 st positiva patchar och 427 st negativbilder med storlekar på 400x400 och 800x800 pixlar. I varje negativbild kan (höjd-30)*(bredd-30) stycken 30x30 pixlar stora patchar plockas för träning. De bilder som används för test är lika många till antalet som de som används för träning.

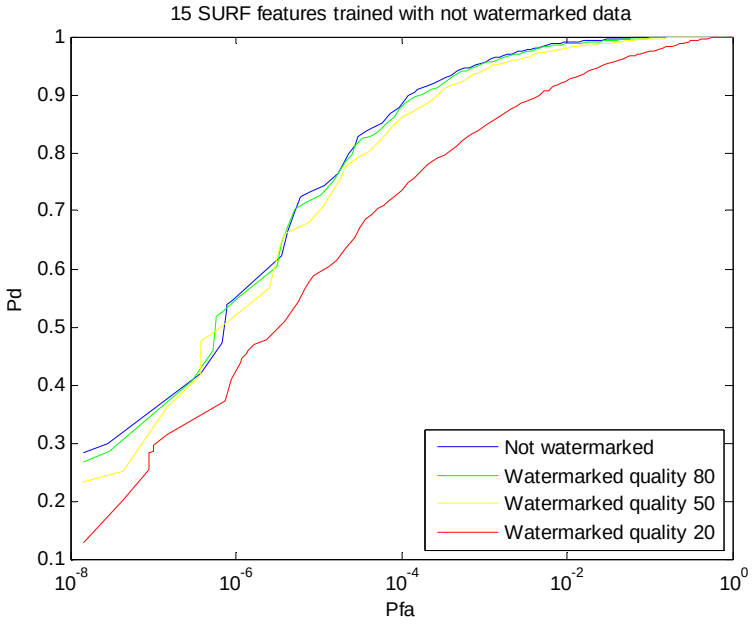
Tre dataset som baseras på det icke vattenmärkta grunddatat genereras med den vattenmärkningsmetod som beskrivs i kapitel 4.4. Ett vattenmärke läggs in i bilderna med olika krav på att klara JPEG-komprimering. Tre olika hanterade nivåer av JPEG-komprimering används: 20, 50 och 80. Totalt finns alltså 4 dataset, de tre med olika vattenmärkningskvalitet och det utan vattenmärkning.

För att testa hur vattenmärkning påverkar metoden har vi tränat detektorn på olika sätt. I det ena fallet används bara icke vattenmärkt data till träning och prestanda testas på vattenmärkt data samt data som inte är vattenmärkt. I det andra fallet tränas detektorn med data som är vattenmärkt kombinerat med data som inte är vattenmärkt (totalt 20700 positiva patchar). Prestanda testas på samma sätt som i första fallet. Även de särdrag som används i boostingdetektorn varieras genom att använda Viola/Jones rektangelsärdrag (Viola, 2001) eller SURF-särdrag (Bay, 2008). Resultatet av testerna finns som ROC-kurvor i figur 6-9.

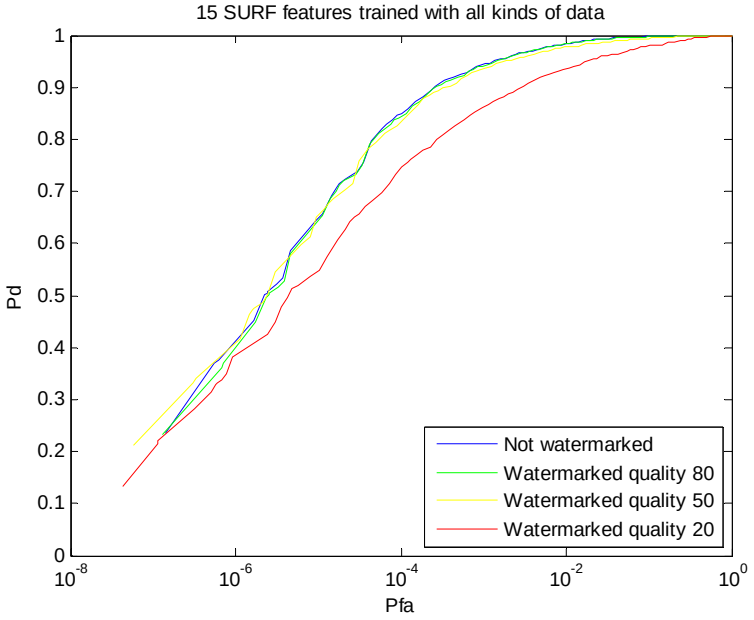
Resultatet visar att vattenmärkning påverkar resultatet, men inte speciellt mycket förrän bilderna förändras kraftigt (kvalitetsnivå 20). En jämförelse mellan de olika SURF-fallen visar att om man tränar med alla typer av data istället för bara icke-vattenmärkt data så blir resultatet för de vattenmärkta bilderna med kvalitet 20 marginellt bättre, men då blir resultatet för icke-vattenmärkt data och lätt vattenmärkt data sämre.

Genom ROC-kurvan ser vi hur stor andel av fordonen som detekteras (Pd) givet en viss falsklarmsannolikhet (Pfa). Vid $Pfa=10^{-4}$ uppnås Pd=0.746, då detektorn tränad med alla typer av data används, jämfört med Pd=0.7353, då detektorn tränad med icke vattenmärkt data används. Det blir tvärtom för resultatet på icke vattenmärkt data och även vattenmärkt data med lite artefakter. Resultatet för bilder som inte är vattenmärkta blir så att vid $Pfa=10^{-4}$ fås Pd=0.849 då detektorn tränad med alla typer av data används, jämfört med Pd=0.881 då detektorn tränad med icke vattenmärkt data används. För vattenmärkt data med kvalitet 80 vid $Pfa=10^{-4}$ fås Pd=0.84 med detektorn tränad med alla typer av data och Pd=0.874 med detektorn tränad med icke vattenmärkt data. Från testerna kan man dra slutsatsen att det inte är bra att lägga in extra träningsdata med dålig kvalitet bara för att utöka sin träningsdatamängd. Det kan bli bättre för att detektera i bilder med dålig kvalitet men prestanda på bilder med bra kvalitet blir lidande.

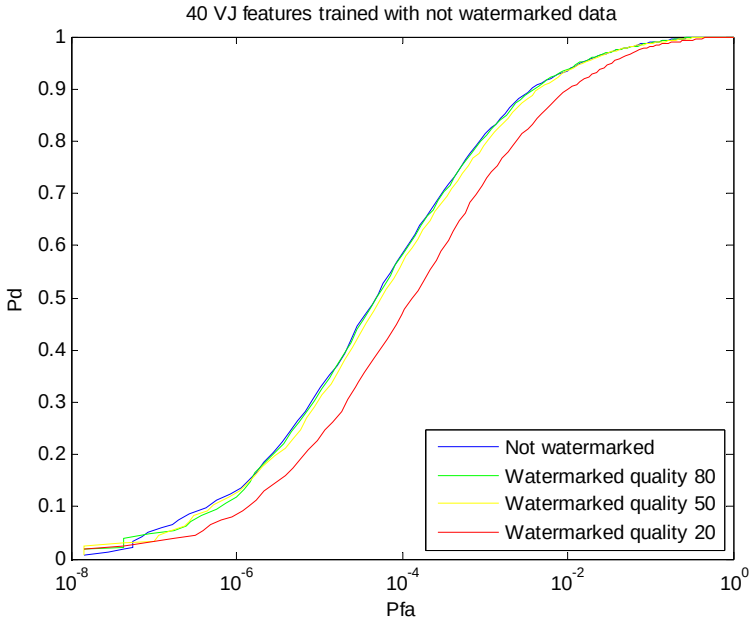
Resultatet för Viola/Jones (VJ) rektangelsärdrag är liknande som de för SURF-särdragen, fast med mindre variation. Det beror troligtvis på att SURF särdragen arbetar med summor i regioner i derivator-bilder medan VJ-särdragen arbetar med summor i regioner i de vanliga bilderna. Vattenmärkningen ger artefakter som producerar kanter vilket förändrar svaret mer för SURF-särdragen än för VJ-särdragen.



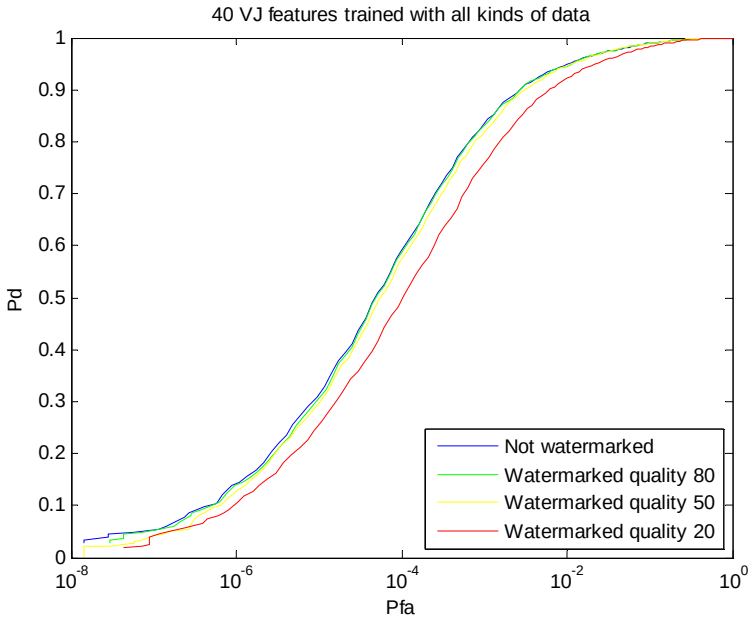
Figur 6: ROC-kurvor för en detektor med 15 SURF-särdrag tränad på icke vattenmärkt data och testad separat på icke vattenmärkt data samt vattenmärkt data med tre olika kvalitetskrav.



Figur 7: ROC-kurvor för en detektor med 15 SURF-särdrag tränad på både vattenmärkt data och icke vattenmärkt data. Den är testad separat på icke vattenmärkt data samt vattenmärkt data med tre olika kvalitetskrav.



Figur 8: ROC-kurvor för en detektor med 40 VJ-särdrag tränad på icke vattenmärkt data och testad separat på icke vattenmärkt data samt vattenmärkt data med tre olika kvalitetskrav.



Figur 9: ROC-kurvor för en detektor med 40 VJ-särdrag tränad på både vattenmärkt data och icke vattenmärkt data. Den är testad separat på icke vattenmärkt data samt vattenmärkt data med tre olika kvalitetskrav.

3 Individuella vattenmärken

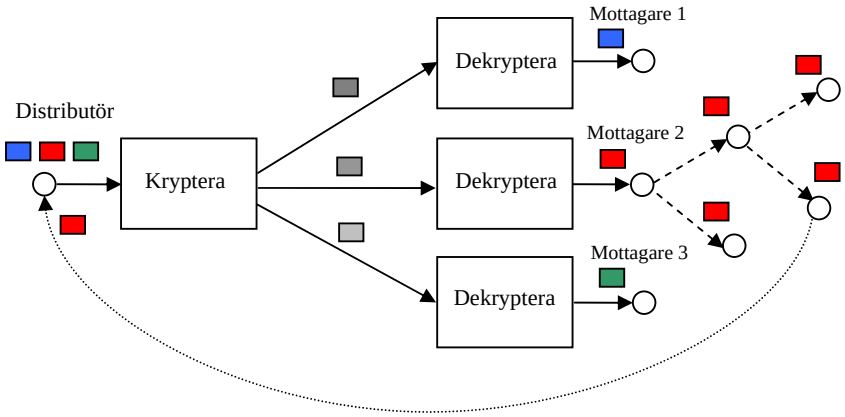
Detta kapitel handlar om vattenmärken som är unika för varje auktoriserad användare. Detta gör att om sensordata sprids otillåtet, går det utgående från den spridda kopian spåra vilken av de betrodda användarna, kopian ursprungligen lämnats ut till.

3.1 Exempel på problemställning

I många tillämpningar kan det vara väsentligt att data bara överförs till eller finns tillgängligt för de avsedda användarna. Denna hantering av hemlig information brukar skötas genom kryptering. Problemet uppstår om någon betrodd användare medvetet eller omedvetet sprider den dekrypterade informationen vidare.

Om sådan hemlig information dyker upp på hemsidor, i media eller når till andra icke-auktoriserade parter måste läckan snabbt tätas. Problemet är att få reda på var läckaget skett. Detta kan vara extra svårt om det endast är vid ett tillfälle som information spridits vidare.

Ett sätt att spåra otillåtet spridd information är att använda individuellt vattenmärk information. Det betyder att varje användares tillgängliga data innehåller unika delar. För bilder, ljud och video kan det handla om ett svagt brus. Även textdokument kan vattenmärkas men det är svårare.



Figur 10: En av de tre betrodda mottagarna sprider vidare bilden efter att den har blivit dekrypterad. Distributören får tag på en sådan spridd bildkopia som innehåller ett individuellt vattenmärke (här representerat genom sin färg). Då kan den betrodda mottagaren som kopian lämnats ut till identifieras, vilket förhoppningsvis gör att den som är skyldig till den otillåtna spridningen kan hittas.

Om det läckta materialet sprids till ett ställe där det är allmänt tillgängligt (exempelvis via en hemsida) går det då undersöka vilket individuellt vattenmärke det innehåller och därmed spåra läckan, se figur 10. Om däremot det läckta materialet inte finns tillgängligt går det inte heller utnyttja vattenmärket. Däremot kan vetskapen om att materialet är individuellt vattenmärkt avskräcka användare från att läcka informationen.

3.2 Individuella vattenmärken och attacker

Individuella vattenmärken[3] är unika för varje betrodd användare av ett system. I fallet att exempelvis bilder överförs från en central till tio olika mottagare skall då bilden som en mottagare får, vara inbäddad med ett vattenmärke som är unikt för honom.

Individuella vattenmärken skall ses som ett komplement till kryptering, där krypteringen skyddar sensordata från yttre attacker under distribution och lagring. Krypteringen skyddar dock inte om klartextversion av sensordata (dekrypterat) sprids vidare avsiktligt eller av misstag. För att det individuella vattenmärket skall komma till användning måste då en sådan olovligt spridd kopia hittas, så att märket kan utvinnas och spåra spridningskällan, se figur 10. Individuella vattenmärken förhindrar inte otillåten spridning fysiskt, men om otillåten spridning skett kan individuella vattenmärken användas för att spåra spridningskällan.

Om en bild innehåller ett unikt vattenmärke kan det ha en avskräckande effekt på en potentiell spridare, eftersom att en sådan spridd kopia av bilden senare kan användas för att spåra läckan. Detta skulle alltså leda tillbaka till den skyldiga användaren. Därför är det troligt att en sådan användare, innan spridningen, attackerar bilden för att försöka radera eller skada det individuella vattenmärket. Individuella vattenmärken måste därför vara robusta mot attacker. För bilder kan attacker bestå av allt från förstörande komprimering (som JPEG), brus, filtrering till storleksförändringar och många andra operationer. För att vara effektivt bör därför ett individuellt vattenmärke kunna motstå många olika typer av attacker.

Om fler än en användare samarbetar för att otillåtet sprida vidare sensordata, så kan de tillsammans attackera informationen. Om två användare samarbetar för att ta bort de både individuella vattenmärken i bilder går det exempelvis ta medelvärdet av de båda bilderna, ta varannan pixel eller mer analysera skillnaderna mellan bilderna för att göra en anpassad attack. Om inte de kan radera sina märken går det ju också försöka likna en tredje (oskyldig) användares vattenmärke. En individuell vattenmärkningsalgoritm bör därför även vara robust mot samarbetande attacker. Detta gäller också när fler än två användare samarbetar om attacken.

Förutom bilder, video och ljud kan även textdokument vattenmärkas. Detta kan bygga på att texten måste vara kvar inom samma format, exempelvis som en pdf-fil. Små förskjutningar av placeringen av bokstäverna utgör då själva vattenmärket. Ett annat sätt att unikt märka texter är att vissa ord byts ut mot synonymer (eller att hela meningar formuleras om). Denna metod är snarare en del av området ”fingerprinting” än ”watermarking”. Det är viktigt att synonymerna väljs med omsorg för att det inte skall kunna uppstå missförstånd över textens innehåll. En fördel med synonymfallet, över föregående metod, är att vattenmärket följer med även om texten flyttas mellan en pdf-fil, txt-fil, e-post eller en hemsida. Däremot går det inte få in någon stor kapacitet i denna form av märkning, utan en lång text behövs för att bädda in många bitar med information.

3.3 Effektiv överföring och lagring

För att kunna använda individuella vattenmärken i ett distributionssystem får inte kostnaden av att införa individuell inbäddning vara större än den nytta den skapar. Kostnaden kan vara i form av:

- Krav på högre bandbredd på grund av de individuella vattenmärkena.
- Lägre distributionshastighet genom inbäddningen.
- Hur flexibelt distributionssystemet är vad gäller utbyggnad och nya användare.
- Monetär kostnad av själva införandet.

[3] Internationellt används ibland uttrycket Forensic watermarks för denna typ av vattenmärken. Även Fingerprinting används men syftar då egentligen på ett närliggande tillämpningsområde.

3.3.1 Överföring via unicast, multicast och broadcast

En del forskning har utförts på att distribuera individuellt märkt data utan en kraftig ökning av bandbredden.

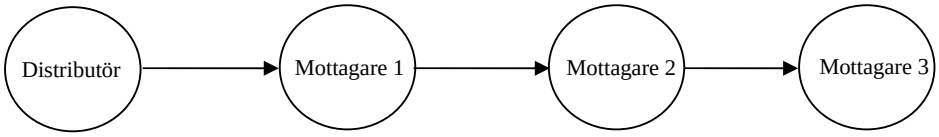
Vi har valt att dela in de olika metoderna för unikt vattenmärkt material i fem kategorier: Källbaserad vattenmärkning, nätverksbaserad vattenmärkning, distribution av unika fragment, distribution av delade fragment och klientbaserad vattenmärkning. Detta har vi tidigare publicerat i (Stenborg, 2009 March) (Stenborg, 2009 December). Kortfattat går de ut på:

- **Källbaserad vattenmärkning**, är inte skalbar utan bygger på att en distributör vattenmärker varje kopia och överför dem en och en till varje mottagare.
- **Nätverksbaserad vattenmärkning**, bygger på att det finns speciella närverksnoder som utför vattenmärkningen (Judge, 2000) (Brown, 1999). Dessa metoder ställer hårda krav på typen av nätverk och kräver lagrad information om vad som har skickats när och till vem.
- **Distribution av unika fragment**, görs genom att en ström delas in i två delar (Wu, 1997) (Luh, 2005). En del av strömmen innehåller inte något vattenmärke och distribueras effektivt via broadcast eller multicast. En andra del strömmen vattenmärks individuellt och överförs sedan till en mottagare var via unicast. Vid ett stort antal mottagare kräver dock den vattenmärkta delen mycket kapacitet. Dessutom måste det på något sätt garanteras att den icke vattenmärkta delen, inte innehåller någon information som har något värde, utan tillgång till den vattenmärkta delen.
- **Distribution av delade fragment**, bygger på att strömmen med information delas in i fragment och dessa finns i två versioner innehållandes olika vattenmärken. En unik uppsättning av delarna från de båda strömmar kan sedan läsas av varje mottagare genom kryptering av innehållet. Genom det blir det sammanställda mottagna innehållet individuellt vattenmärkt (Parviainen, 2001) (Zhao, 2006). En nackdel med dessa metoder är att de kräver högre överföringskapacitet jämfört mot icke vattenmärkt material. Dessutom måste fragmenten delas upp på sådant sätt att de går att robust vattenmärka var för sig.
- **Klientbaserad vattenmärkning**, ger oftast effektiva metoder och bygger på att det överförda materialet förvrängs hos distributören. Samma förvrängda material överförs till alla mottagare. Varje mottagare måste sedan genom en unik nyckel återställa ursprungliga delar ur de förvrängda, men nyckeln är satt sådan att en viss unik distorsion införs. Varje mottagare får därigenom ett unikt vattenmärkt material (Anderson, 1997) (Emmanuell, 2003) (Kundur, 2004) (Stenborg, 2009 March).

Den första kategorin bygger alltså på att materialet överförs via unicast med en ström till varje användare. Detta är inte alls bandbreddseffektivt och kräver dessutom mycket beräkningskapacitet hos distributören eftersom alla vattenmärken måste bäddas in där före överföringen.

Metoder i de övriga fyra kategorierna är mer anpassade för samtidig överföring via multicast eller broadcast, vilket är mer bandbreddseffektivt. *Distribution av unika fragment* kräver en hel del beräkningskraft hos distributören, medan metoderna i de övriga tre kategorierna är mindre beräkningsbelastande för distributören eftersom själva inbäddningen av det mottagarunika märket sker i nätverket eller vid mottagarens klient.

Två av dessa kategorier (*Distribution av delade fragment* och *Klientbaserad vattenmärkning*) är dessutom väl anpassade för situationer där informationen finns lagras i en central som användarna får koppla upp sig mot när de vill få tillgång till någon information. Detta gäller alltså inte överföring till många användare i realtid, utan individuell åtkomst av äldre information.

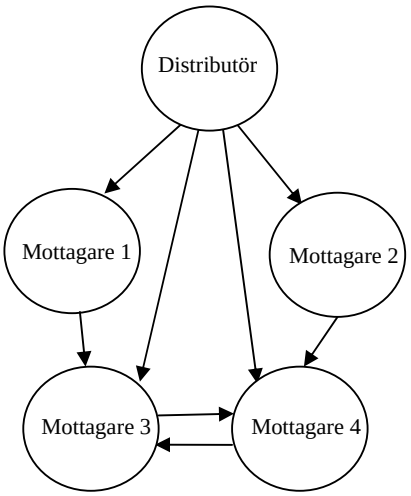


Figur 11: En linjär återdistributionskedja där användare 1 distribuerar vidare till användare 2 och så vidare.

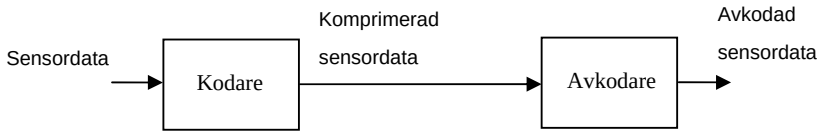
3.3.2 Överföring via peer-to-peer och ad hoc-nätverk

Under projektets andra halva har vi mer och mer intresserat oss för hur individuellt vattenmärkt information skall kunna överföras via nyare typer av nätverk som peer-to-peer baserade eller ad hoc-nätverk. Detta är då närbesläktat med problematiken med auktoriserad återdistribution som togs upp i (Stenborg, 2009 December). Med auktoriserad återdistribution syftar vi på att en auktoriserad mottagare skall kunna dela med sig av sin information till andra auktoriserade mottagare, exempelvis kolleger. Problemet är då att materialet fortfarande bör kunna spåras till den sista auktoriserade användaren om det läcker ut till icke auktoriserade mottagare. I figur 11 illustreras problematiken. Detta kan ses som ett fall av ad hoc-nätverk där Mottagare 1 och 2 agerar som noder för att distributörens material skall nå mottagare 3. Den information som mottagare 3 får måste alltså innehålla ett individuellt vattenmärke som är kopplat till honom, inte till mottagare 1 eller 2. Samtidigt skall mottagare 1 inte få tillgång till individuella vattenmärket som är kopplat till mottagare 2 och 3.

En annan typ av överföring är den som sker genom peer-to-peer nätverk (förkortat P2P). Det finns många typer av sådana nätverk men vi tänker oss en version med en distributör och flertalet användare, figur 12. Detta påminner exempelvis om den form av P2P som används av musiktjänsten Spotify. På liknande sätt som för återdistribution och ad hoc-nätverk måste då distribution av information med individuella vattenmärken lösas för P2P-fallet.



Figur 12: Ett peer-to-peer nätverk med en distributör och fyra användare. Exempelvis kan Mottagare 1 få information överfört från Distributören. Om därefter Mottagare 3 vill få tillgång till samma information kan det endera ske via Distributören eller Mottagare 1. Om det för nätverket är enklare att överföra från Mottagare 1 kommer strömmen att komma därifrån.



Figur 13: Sensordata komprimeras i en kodare. För att kunna läsa det komprimerade sensordatat måste den packas upp i en avkodare.

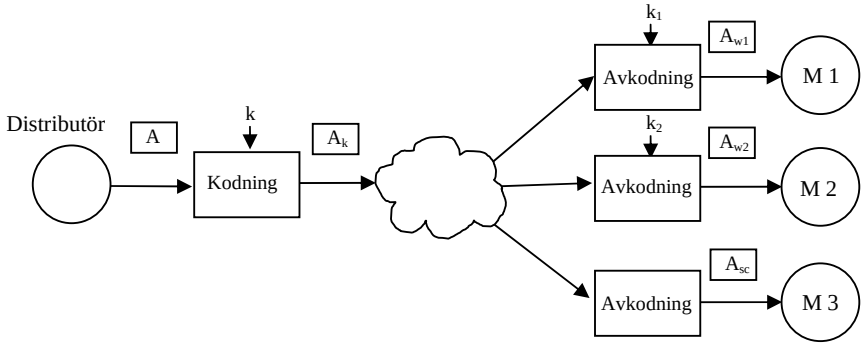
Inom vårt arbete (Stenborg, 2010) (Stenborg, 2011) visar vi några sätt att utföra distribution av individuellt vattenmärkt material i ett peer-to-peer nätverk. Det är en metod (Parviainen, 2001) inom kategorin *Distribution av gemensamma fragment*, samt vår egen MTW-metod (Stenborg, 2009 Mars) inom kategorin *Klienbaserad vattenmärkning* som vi använt oss av. Den första metoden ger upphov till ett flera intressanta frågeställningar men leder till mer komplexa överföringsförfaranden och kräver oftast dubbla bandbredd och lagringsutrymmet jämfört med icke vattenmärkt material. Vi anser att den andra metoden, vår så kallade Modified Transform Watermarking (MTW) algoritm, är den som är bäst anpassad för P2P-nätverk. Den kan dessutom användas i många andra typer av distribution. Denna metod kan till och med användas i den form av återdistribution som beskrevs i figur 11. Nästa avsnitt tar upp denna metod.

3.3.3 Modified Transform Watermarking

Vår egenutvecklade klientbaserade vattenmärkningsmetod kan användas på stillbilder och video. Arbetet med Modified Transform Watermarking har tidigare beskrivits i (Stenborg, 2009 March) (Stenborg, 2009 October) och (Stenborg, 2009 December).

Metoden använder sig av DCT-transformen vid vattenmärkningen. DCT-transformen är vanligt förekommande vid komprimering av video och stillbilder (som JPEG, MPEG-2, AVC osv), figur 13. Istället för att använda den vanliga DCT-transformen förändras elementen i transformen genom några operationer. Denna nya transform används sedan vid komprimeringen. För en stillbild medför detta att om bilden sedan avkodas genom den vanliga inverterade DCT-transformen kommer resultatet bli en kraftigt skadad bild, figur 14. Istället måste en speciell användarunik transform användas vid avkodningen, genom att en individuell avkodningsnyckel. Den transformen är liknande men inte exakt motsvarande den inverterade transformen som användes i kodaren. Detta medför att bilden nästan blir återskapad men innehåller dels komprimeringsdistorsion och dessutom en unik distorsion som ges av skillnaden mellan transformerna i kodare och avkodare. Denna användarunika distorsion är det individuella vattenmärket.

En stor fördel med MTW och de andra klientbaserade vattenmärkningsmetoderna är att mängden data som överförs i nätet är betydligt lägre än för de andra kategorierna av effektiv distribution. Detta gör att de klientbaserade metoderna är mycket bättre anpassade för stora grupper av användare. *Klientbaserad vattenmärkning* är dessutom anpassad för system där ett antal användare kan fråga efter lagrat sensordata. Sensordata ligger då lagrat på samma format som under överföring. Användaren måste då använda sin unika nyckel för att använda sensordata och bäddar då in sitt individuella vattenmärke, precis som när den mottagits efter överföring.

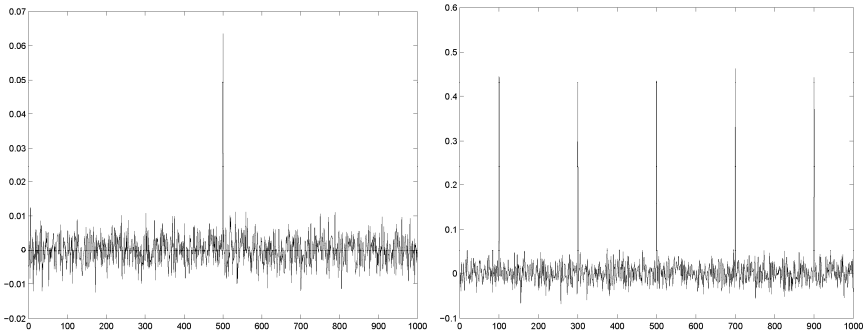


Figur 14: Klientbaserad vattenmärkning. Modified Transform Watermarking använder en hemlig nyckel *k* under kodningen av bilden som vrider till transformerna. Denna version av bilden överförs sedan. Detta gör att om en mottagare (*M3*) inte har någon nyckel kommer den avkodade bilden A_{sc} vara kraftigt skadad och kan ej användas. De betrodda mottagarna (*M1* och *M2*) får en varsin unik nyckel (k_1 och k_2) som används vid avkodningen. Dessa vrider nästan tillbaka kodnings-transformerna, men lämnar ett litet unikt brus som utgör det individuella vattenmärket i A_{w1} och A_{w2}.

I den första versionen av MTW går det inte direkt utvinna vattenmärket från en stillbild. Istället måste en bild som skall undersökas korreleras med alla individuellt vattenmärkta bilder, för att se vilket märke den innehåller. Det gör att bara distributören eller någon annan pålitlig instans kan utföra vattenmärkningskontrollen. Dessutom är det omständligt att genomföra en sådan undersökning, speciellt om antalet individuella vattenmärken är stort. Fast om vattenmärken inte skall kontrolleras så ofta kan det ändå vara möjligt att använda denna form av vattenmärkningsextraktion.

Vi har utarbetat ett par nya versioner av MTW där vattenmärket kan utvinnas mer lätthanterligt utan att korrelation med alla bilder behövs. Ingen av dessa versioner har dock ännu implementerats.

Den första versionen av MTW har genomgått ett antal tester där individuellt vattenmärkta stillbilder (svartvita med 8 bitar per pixel) har attackerats för att försöka radera vattenmärket. En attack där rätt starkt brus med medelvärde 0 och standardavvikelse 20 lagts på en bild, ger enligt figur 15 (a) ändå en tydlig indikation på vilket av de 1000 vattenmärkena det är som bilden innehåller. Om fem användare samarbetar och tar medelvärdet av sina olika bilder kommer alla fem tydligt synas vid korrelation, vilket visas i figur 15 (b). MTW är därför robust mot dessa båda specifika attacker. Vi har även testat andra attacker (både singelattacker och samarbetande attacker) och uppnått liknande resultat.



(a) (b)

Figur 15: (a) En användare har attackerat en stillbild genom att lägga på ett brus med standardavvikelse 20 och sedan spridit den vidare. Genom att korrelera resultatet från de 1000 användarna och den attackerade bilden framgår det klart att det är användare nr 500 som är källan till läckan. (b) Fem användare samarbetar och tar medelvärdet av sina fem stillbilder för att försöka tvätta bort vattenmärket. Korrelation mellan den attackerade bilden och de 1000 användarnas unika bilder ger att användare 100, 300, 500, 700 och 900 är de skyldiga.

4 Förändringsindikation

Detta kapitel handlar om integritet av sensordata. Bräckliga vattenmärken kan användas för att upptäcka om sensordata har blivit manipulerad eller skadad. Det finns möjlighet att se var förändringen skett samt möjligheter att delvis återskapa skadade områden.

4.1 Exempel på problemställning

I många tillämpningar är det väsentligt att kunna avgöra om en information är oförändrad. Detta gäller även för sensordata. För fotografier gick det tidigare att gå tillbaka till filmnegativet för att se hur bilden ursprungligen såg ut. För digitala bilder finns det inget sådant negativ som innehåller ”sanningen”.

En lägesbild över ett område har införskaffats genom nytagna flygbilder. Det finns risk för att det någonstans i överföringskedjan har uppkommit skador eller avsiktlig manipulering av denna sensordata. Därför är det viktigt att med säkerhet kunna avgöra att den mottagna informationen stämmer med ursprungliga data.

Ett annat scenario kan vara övervakningsvideo runt en strategiskt viktig byggnad. Även där gäller det att försäkra sig att både ny och lagrad videodata stämmer. Ett flertal människor har tillgång till denna data och skulle därmed kunna manipulera informationen genom bildbehandlingsprogram som Photoshop, GIMP och After Effects.

Denna frågeställning om huruvida sensordata är oförändrad eller ej, kan lösas genom att använda digitala signaturer och/eller bräckliga vattenmärken. På det sättet går det upptäcka om någon information inte är pålitlig. För stillbilder finns det möjlighet att se inom vilket område som en bild har blivit förändrad. Detta kan vara till hjälp för att avgöra hur allvarlig förändringen är och vilken typ av information som kan saknas. Det finns dessutom möjlighet att bädda in ett robust vattenmärke som innehåller återskapande information vilket gör att skadade områden delvis kan rekonstrueras.

4.2 Digitala signaturer och bräckliga vattenmärken

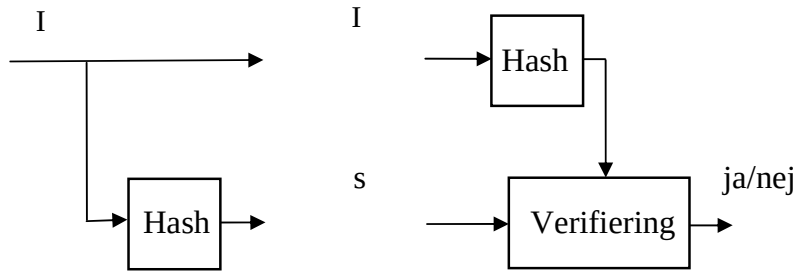
Det finns flera sätt att bekräfta att sensordata är oförändrad. Den mer kryptografiska metoden med Digitala signaturer eller olika former av bräckliga vattenmärken (fragile watermarks) kan användas. Bräckliga vattenmärken kan delas in i dem som bygger på en bräcklig inbäddningsmetod och dem som bäddar in någon form av digital signatur.

4.2.1 Digitala signaturer

Digitala signaturer kan användas för att autentisera digitala dokument. En bra genomgång av området med inriktning mot multimediafiler beskrivs av Sun (2005).

Här tar vi upp autentisering av dokument som skickas från en sändare A till en mottagare B. Förenklat går det till så att ett hash-värde (Rivest, 1992) beräknas på delar av eller hela det digitala dokumentet hos A. Hash-värdet är den digitala signaturen och den skickas tillsammans med det digitala dokumentet till B. När B har mottagit det digitala dokumentet använder han samma hash-funktion och jämför svaret med den signatur som A skickade, se figur 16. Om de stämmer kan B med stor sannolikhet förlita sig på att det digitala dokumentet inte har blivit förändrat under distributionen. Själva hashningen måste göras med privata respektive öppna nycklar för att ingen tredje part som vill förändra dokumentet skall kunna anpassa signaturen till förändringen.

Digitala signaturer kan även användas i andra sammanhang än överföring, exempelvis för att försäkra sig om att ett lagrat dokument inte har blivit förändrat. Det görs genom att det förutom det lagrade dokumentet finns en autentiseringssignatur lagrad. Den används återigen för att jämföras med resultatet från hashningen av dokumentet.



Figur 16: Digital signatur. Förenklad beskrivning av hur en signatur *s* skapas genom en hash av en bild *I*. Vid ett senare tillfälle används signaturen för att verifiera bilden genom jämförelse med en ny hash av bilden.

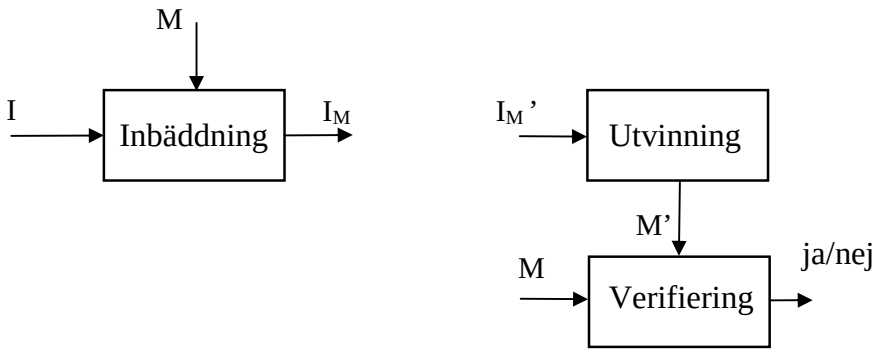
Ett problem med denna typ av förändringindikation är att signaturen och det digitala dokumentet måste hållas samman. Om det digitala dokumentet exempelvis är en bild så måste signaturen alltid finnas med. Om signaturen då lagras i en header till bilden finns risk för att den försvinner om bilden genomgår någon formatförändring. Skulle en bild spridas och hamna på en internetsida är risken stor att den har bytt format och kanske dessutom komprimerats. Hur skall vi då kunna avgöra om det som visas på bilden stämmer eller om den har blivit manipulerad?

4.2.2 Bräckligt inbäddade vattenmärken

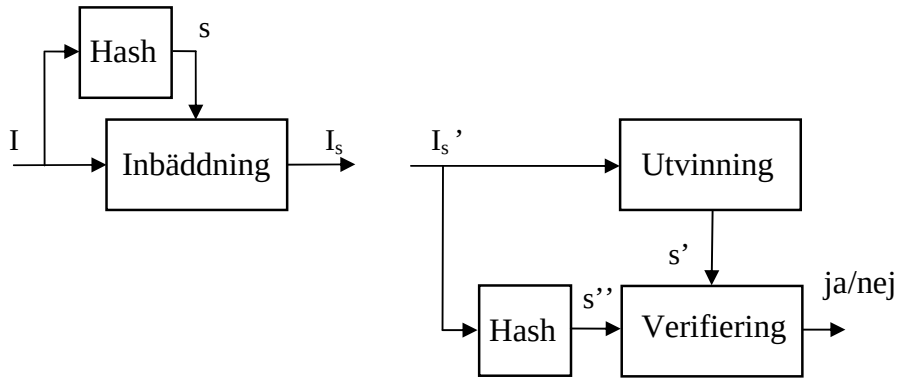
Ett annat sätt att autentisera sensordata är att använda en bräcklig vattenmärkningsmetod. Det betyder att vattenmärket kommer bli skadat om de sensordata den är inbäddad i blir skadade.

Genom att mottagaren i förväg vet vilket bräckligt vattenmärke som har bäddats in går det att utvinna märket ur värdbilden och jämföra de båda märkena, figur 17. Är det utvunna märket det samma som det inbäddade kan mottagaren vara rätt säker på att de sensordata som innehållit märket inte har påverkats.

När det bräckliga vattenmärket bäddas in så förändras bilden något. Vid vissa tillämpningar som för medicinska bilder ses detta som ett allvarligt problem. I de flesta tillämpningar är denna påverkan dock minimal. För vissa algoritmer finns det möjlighet att efter verifiering tvätta bort det bräckliga vattenmärket. Det går dock aldrig garantera att denna borttvättning inte inför vissa förändringar så den tvättade bilden skiljer sig något från originalbilden (Cox, 2008).



Figur 17: Förändringsindikation genom att ett för mottagaren känt vattenmärke *M* bäddas in bräckligt eller halvbräckligt hos sändaren. Mottagaren kan sedan utvinna märket *M´*ur bilden och jämföra det mot märket *M* för att se om bilden har blivit förändrad.



Figur 18: Förändringsindikation genom bräcklig eller halvbräcklig signatur. En signatur *s* beräknas på bilden genom exempelvis en hashfunktion. Denna signatur bäddas in i bilden hos sändaren så I_s bildas. Mottagaren av en bild I_s' kan utvinna det inbäddade vattenmärket *s´*och även använda hashfunktionen för att beräkna bildens signatur *s´´*. Om *s'* och *s''* inte är lika har bilden blivit påverkad. Denna beskrivning är förenklad eftersom inga nycklar används för att beräkna signaturen eller utvinna vattenmärket.

4.2.3 Digitala signaturer som vattenmärken

En annan typ av bräcklig vattenmärkning får vi om själva vattenmärket beror på bilden. Om vattenmärket i själva verket är den digitala signaturen (avsnitt 4.2.1) behöver inte det bräckliga vattenmärket vara känt i förväg för en mottagare.

Som vi ser i figur 18 beräknar vi först en bräcklig signatur på bilden och sedan bäddar vi in den i bilden. Då har bilden blivit påverkad och om en ny signatur beräknas på den inbäddade bilden bör den inte vara den samma som den första signaturen. För att lösa detta problem kan inte signaturen beräknas på alla bildens bitar utan vissa bitar (eller delar) av bilden används för att beräkna signaturen och andra bitar (eller delar) av bilden används till inbäddningen av vattenmärket. På så vis kommer signaturen som beräknas på den inbäddade bilden vara den samma som signaturen som beräknades först, förutsatt att ingen annan förändring av bilden har skett.

Vid utvinningen av det inbäddade märket får vi enligt figur 18 ut märket/signaturen *s'*. Om själva inbäddningsmetoden är robust kommer då $s' = s$ även om bilden har utsatts för förändringar. Eftersom signaturberäkningen är bräcklig kommer dock $s'' \neq s$. Detta leder till att i verifieringssteget kommer vi att finna att $s' \neq s''$ och vi vet därmed att bilden har blivit förändrad.

4.3 Autentisering

Det finns både halv-bräckliga och helt bräckliga vattenmärkningsalgoritmer. Autentiseringen kan gälla hela sensordata eller indikera vilka delar som är skadade. Dessutom finns det möjlighet att delvis återskapa förändrade områden.

4.3.1 Halvbräckliga och bräckliga vattenmärken

För tillämpningen förändringsindikation kan själva märket som bäddas in vara av olika typer, endera är det ett fast märke eller också en signatur av något slag som är beroende av originalbilden. Nedan beskriver vi dessa typer av märken. Notera att vi nu bara diskuterar själva märket innan inbäddningen och inte inbäddningsmetoden.

- **Känt vattenmärke** betyder att vattenmärket som bäddas in redan är känt av mottagaren och att det märket inte är beroende av innehållet i bilden. Detta beskrivs i 4.2.2.
- **Bräcklig signatur** är en typ av märke som beror på bilden på sådant vis att om bilden förändras lite grann kommer den beroende signaturen att förändras mycket. Att vi kallar denna typ av märken för signaturer beror på likheten med digitala signaturer. Detta beskrivs i 4.2.3.
- **Halvbräckliga signaturer** är också ett märke som är beroende av bilden fast till skillnad från den bräckliga signaturen i föregående punkt förblir denna typ av signatur den samma även om bilden genomgår mindre förändringar som exempelvis måttligt förstörande komprimering. Om däremot bilden genomgår större förändringar, som att delar raderas eller byts ut, kommer den halvbräckliga signaturen som beräknas utgående från den påverkas.

Beroende på vilka bitar eller delar av bilden som vi väljer för att beräkna signaturen kan vi styra om den skall vara bräcklig eller halvbräcklig. Om vi till exempel beräknar signaturen på frekvenskomponenter som är invarianta mot JPEG-komprimering kommer den bli halvbräcklig och robust mot just den typen av komprimering. Se figur 19 för exempel.

(a) (b)

(c) (d)

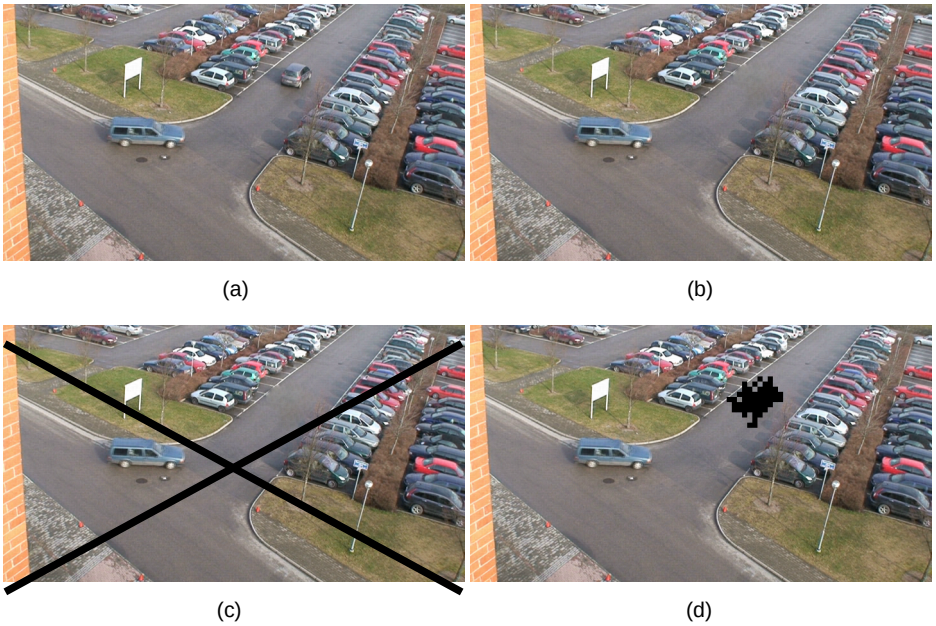
Figur 19: Exempel över bild som har vattenmärkts med en halvbräcklig signatur. (a) Vattenmärkt bild med halvbräcklig signatur som skall klara en JPEG-komprimering med kvalitet 50. (b) Föregående bild JPEG-komprimerad med kvalitet 60. (c) Föregående bild manipulerad så att den grå bilen inte längre syns. (d) Förändringsindikation på föregående bild där svarta områden har blivit förändrade.

Signatur\Märkning	Bräcklig	Halvbräcklig	Robust
Bräcklig	B	B	B
Halvbräcklig	B	HB	HB
Känd=Robust	B	HB	R

Tabell 1: Tabell över de olika bräcklighetsfallen. Det är nio olika fall som kan uppstå beroende på val av signatur och inbäddningsmetod. Den av signaturberäkningsmetoden och märkningen som är mest bräcklig styr hur bräcklig hela vattenmärkningsmetoden blir. B = Bräcklig, HB = Halvbräcklig och R = Robust.

Om inbäddningsmetoden av märket är bräcklig eller halvbräcklig, och bilden har blivit förändrad, uppstår situationen att det utvunna märket är $s' \neq s$ i figur 18. Vi får dessutom att den beräknade signaturen är $s'' \neq s$. Med väldigt hög sannolikhet kommer vi då åter igen få att $s' \neq s''$ i verifieringssteget. Med andra ord kan vi välja om vi vill använda en robust, halvbräcklig eller bräcklig inbäddningsmetod.

Eftersom det alltså både finns bräckliga, halvbräckliga samt robusta vattenmärkningsalgoritmer, och bräckliga, halvbräckliga samt kända signaturer/märken som bäddas in, uppstår nio olika möjliga fall för en vattenmärkningsmetod vilket illustreras i tabell 1. Den totala robustheten för en metod blir den lägsta robusthetsnivån av vattenmärkningen och signaturen. Därmed uppstår fem fall med bräcklig vattenmärkning, tre halvbräckliga fall samt ett fall som är helt robust (och därmed inte intressant ur förändringsindikeringssynpunkt). Innan val av metod bestäms bör informationen i denna tabell tas i beaktning så att metoden uppfyller de krav som ställs gällande typ av bräcklighet.



(a) (b)

(c) (d)

Figur 20: (a) Vattenmärkt bild. (b) Manipulerad bild där en bil har tagits bort. (c) Global förändringsindikation kan bara notera att bilden är skadad. (d) Lokal förändringsindikation noterar var bilden är skadad, i detta fall i bildblock om *8 x 8* pixlar markerade med svart. I ett fåtal av blocken har förändringen inte upptäckts, på grund av att förändringen är för liten.

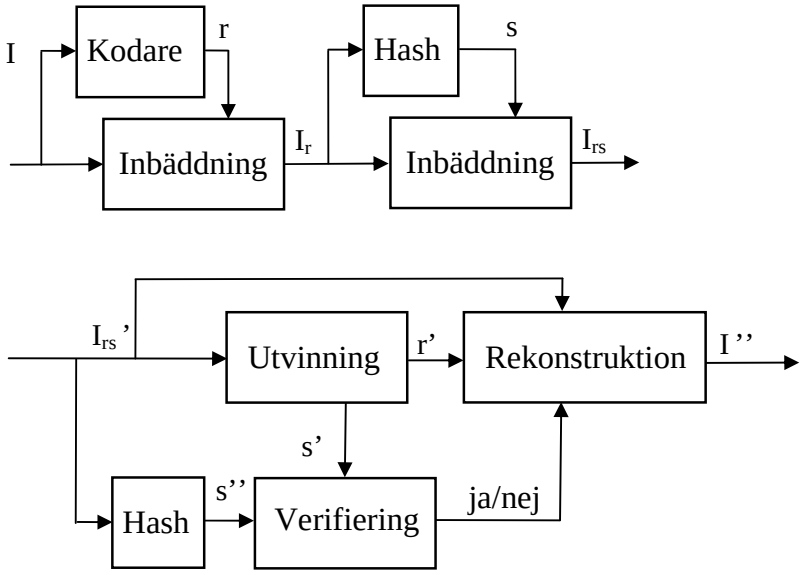
4.3.2 Autentiseringsupplösning

Lite beroende på vilken typ av signatur samt vattenmärkningsmetod som man använder så kan man få olika nivåer av autentiseringsupplösning.

- **Global autentisering** ger bara svar på frågan om bilden har blivit manipulerad eller ej, figur 20 (c).

- **Lokal autentisering** kan däremot avgöra inom vilka områden som en bild har blivit manipulerad, figur 20 (d). Vanligtvis delas bilden in i block och varje sådant block verifieras för sig själv. Det finns dessutom metoder som arbetar på pixelnivå och alltså kan påvisa exakt vilka pixlar som har ändrats. Metoder som ger pixelprecision brukar dock ha problem med att vara helt säkra mot attacker och vanligtvis kommer förändrade pixlar bara upptäckas med en viss sannolikhet.

- **Skadetypsupptäckande autentisering** är en typ av metoder där vattenmärkes-metoden ger en uppfattning om inte bara var utan dessutom vilken typ av attack som har utförts och kanske hur mycket som pixlar har förändrats. Denna typ av märkning kallad ”Telltale watermarking” (Kundur, 1999) har det ännu inte genomförts så mycket forskning på. Det är dock av stort intresse att få fram praktiska vattenmärkningsmetoder som ger denna typ av information.



Figur 21: Återskapande av förstörda delar av bilden genom halvbräcklig vattenmärkning med signatur och återskapningsbild. Bild *I* kodas till en starkt nedsamplad och komprimerad JPEG-bild *r*. Denna bäddas robust in i bilden och bildar I_r på vilken i sin tur en signatur *s* beräknas. Denna signatur bäddas ytterliggare in i bilden och bildar då I_{rs}. Mottagaren av en bild I_{rs}' kan då utvinna signatur *s’* och jämföra den med den beräknade signaturen *s’’*. Genom att också utvinna återskapningsbilden *r’* kan då bilden rekonstrueras i de områden som $s'' \neq s'$. Denna beskrivning är förenklad eftersom inga nycklar används för att beräkna signaturen eller utvinna vattenmärket.

4.3.3 Återskapande av förändrat område

Återskapande (recovery) betyder att det från en skadad bild går att återskapa originalet eller i alla fall få en uppfattning om hur de skadade partierna såg ut ursprungligen.

Vissa förändringsindikeringsmetoder har med ett återskapande steg där de delar av bilden som blivit förändrade delvis kan återställas. Det finns flera metoder för återskapande vattenmärkning men de vanligaste metoderna använder sig av två vattenmärken.

- Som vi ser i figur 21 så skapas först en nedsamplad, hårt kvantiserad och komprimerad version av originalbilden. Denna lågupplösta version av ursprungsbilden bäddas in som ett robust vattenmärke i ursprungsbilden.
- Bilden som skapades i förra punkten blir sedan inbäddad med ett förändrings-indikerande vattenmärke genom någon bräcklig eller halvbräcklig metod.

Vid kontroll av om bilden har förändrats sker då följande två steg:

- Först utvinns det bräckliga eller halvbräckliga vattenmärket och en kontroll sker om bilden har förändrats och i så fall i vilket område av bilden.
- Förutsatt att metoden i föregående punkt kommit fram till att bilden blivit förändrad utvinns därefter den robust inbäddade lågupplösta versionen av originalbilden. Denna används får att ersätta de delar av bilden som blivit markerade som förändrade, figur 21.

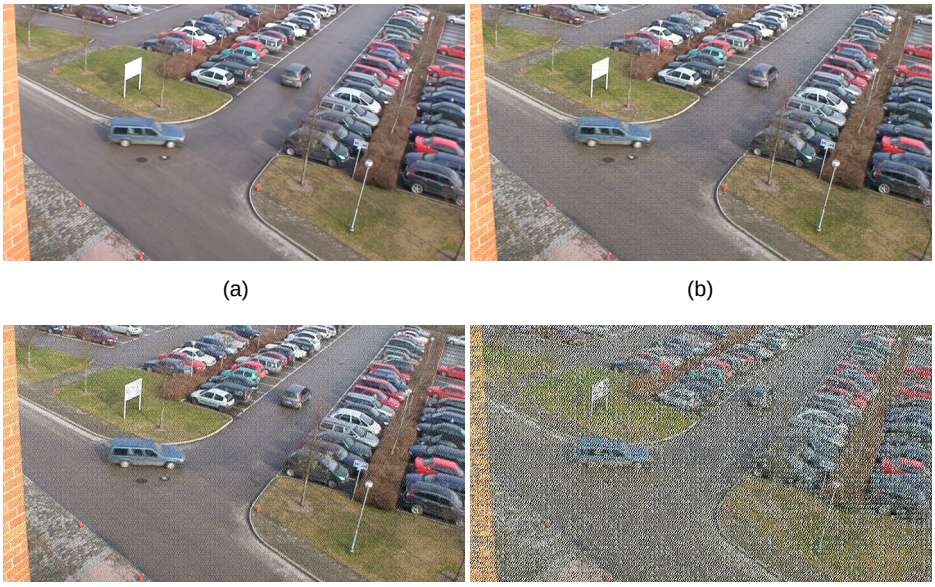
De områden i bilden som återskapas med hjälp av den lågupplösta versionen av originalbilden kommer då inte innehålla alla detaljer som fanns med i originalbilden, men kan förhoppningsvis ge en uppfattning om vad som försvunnit. Exempel på detta kan ses i figur 22**.**

En sak att tänka på är att ett område av den inbäddade lågupplösta versionen av bilden skall bäddas in i ett annat område av bilden än den som den representerar. Exempelvis kan informationen från det övre högra hörnet bäddas in i det nedre vänstra hörnet. Idealt bör dessutom den robusta inbäddningen innehålla redundans så att all vattenmärkes-information finns lagrad på två eller fler ställen i bilden.

(a) (b)

(c) (d)

Figur 22: Exempel på inbäddning och rekonstruktion med återskapningsbild. (a) Ursprungsbilden. (b) Återskapningsbild som är mindre och hårt komprimerad. (c) Vattenmärkt bild som har manipulerats. (d) Rekonstruerad bild där skadade block har ersatts med information från återskapningsbilden.



(a) (b)

(c) (d)

Figur 23: (a) Originalbild. (b) Vattenmärkt bild som klarar JPEG kompression på 25. (c) Vattenmärkt bild med signatur som klarar JPEG-komprimering på 25 och återskapningsbild som klarar JPEG-komprimering på 75. (d) Vattenmärkt bild med signatur och återskapningsbild som klarar JPEG-komprimering på 25.

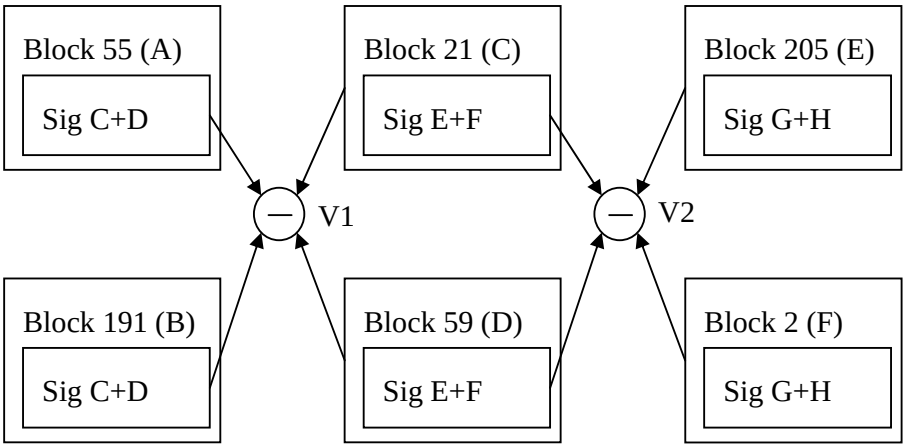
4.4 Implementering av förändringsindikator

Som demonstration av vattenmärkning så har en utvald vattenmärkning/autentiserings-algoritm (Lin, 2000) utvecklats. Denna algoritm arbetar på frekvensplanet på de 8x8 pixel stora block som fås efter att ha DCT-transformerat en bild. Par av block används för att generera signaturbitar från utvalda DCT-komponenter. Då differensen mellan DCT komponenterna används till signaturen så blir algoritmen invariant mot JPEG-komprimering. Signaturbitarna bäddas sedan in i andra DCT block än de som används för att generera signaturen, se figur 24. Relationen mellan vilka blockpar som signaturen beräknas och sedan vilka block de bäddas in i kan med fördel hållas hemlig, till exempel beräknad med en ”nyckel”. Detta för att vara robust mot ”inklistringsattacker” (Holliman, 1997) som går ut på att man känner till blockstorlek och klipper och klistrar block för att dölja förändringar. Eftersom blocken inte är oberoende av varandra i den implementerade algoritmen fungerar inte sådana attacker.

Vid inbäddningen används samma kvantifieringstabeller som vid standard JPEG-komprimering. Detta medför att även inbäddningen är invariant mot viss JPEG-komprimering, alltså en halvbräcklig inbäddning. Distorsionen i den vattenmärka bilden ökar dock beroende på hur hård kvantifiering som ska användas. Detta begränsas genom att använda den kvantifieringstabell som motsvarar den starkaste JPEG-komprimering som vattenmärkningen ska överleva.

På samma sätt som med vattenmärkningen av signaturer så kan en återskapningsbild bäddas in i bilden (se avsnitt 4.3.3). Mängden bitar som behövs för att bädda in återskapningsbilden är oftast många fler än signaturen. Man kan då med fördel använda en kvantifieringstabell som hanterar svagare JPEG-komprimering än signaturen. Figur 23 visar vattenmärkningar som klarar olika grader av JPEG-komprimering.

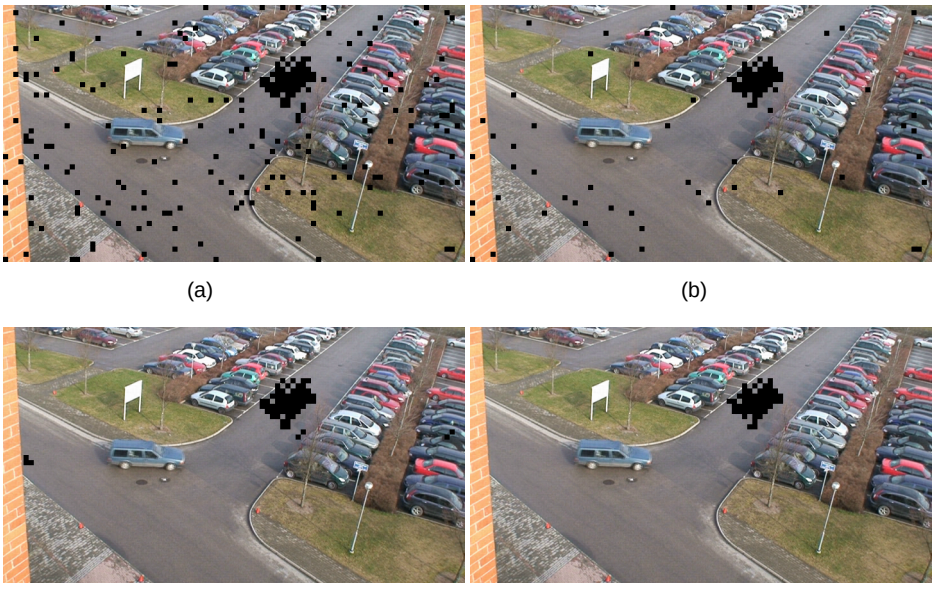
Eftersom algoritmen beräknar signatur på par av block så går det inte att utläsa vilket av de två blocken som är påverkade. Det går heller inte att avgöra om det är signaturblocken eller de block som signaturen bäddas in i som är påverkade. Två metoder har provats för att komma runt detta problem.



Figur 24: Par av block och dess beroende mellan inbäddade signaturer och signaturberäkning. Signaturen för block C och D ligger inbäddat i block A och B samtidigt som signaturen för block E och F ligger inbäddat i block C och D.

Den första metoden är att för varje block där signaturen inte stämmer också kontrollera den signatur som är inbäddad i det blocket. Som visas i figur 24 så är signaturen för block C och D inbäddat i block A och B samtidigt som signaturen för block E och F är inbäddat i block C och D. Om verifieringssteg V1 är negativt men verifieringssteg V2 är positivt, så är det stor sannolikhet att block A och B är påverkade men inte block C och D, vilka då kan uteslutas. På samma sätt: om verifieringssteg V1 är positivt men verifieringssteg V2 negativt, så är sannolikheten stor att block E eller F är påverkade, men inte C och D.

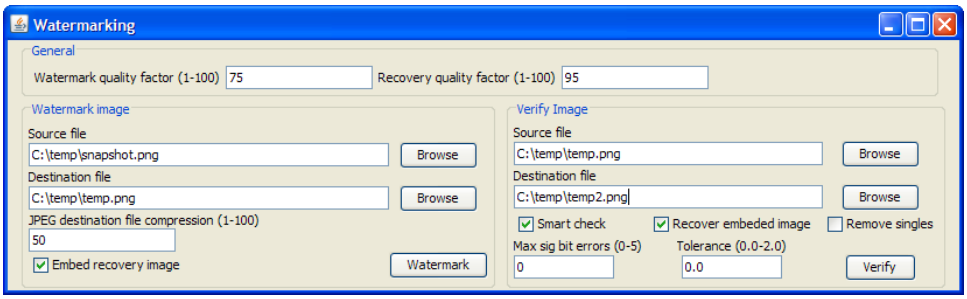
Den andra metoden är att utesluta block som ligger för sig själv. Detta eftersom oftast i en modifierad bild så är ändringen centrerad kring ett visst område. Figur 25 visar exempel på där de olika metoderna har använts, se figur 20 för originalbild och förändrad bild.



(a) (b)

(c) (d)

Figur 25: Exempel på metoder för att bestämma vilka blockpar som har påverkats. (a) Ingen metod påslagen. (b) Första metoden påslagen. (c) Andra metoden påslagen. (d) Både första och andra metoden påslagen.



Figur 26: GUI för vattenmärkningsalgoritm.

Vi har skapat ett GUI för att enkelt demonstrera vattenmärkningen (se figur 26). I detta kan man bland annat ställa in vilken kompressionsnivå i JPEG som vattenmärkningen av signatur och/eller återskapningsbilden ska kunna klara.

Denna implementering av en förändringsindikationsmetod har använts vid skapandet av resultaten som redovisas i figur 19, 20, 22, 23 och 25. För själva manipuleringen då den grå bilen togs bort användes ett vanligt bildbehandlingsprogram.

5 Slutsatser

Inom projektet *Vattenmärkning av sensordata* har tre olika användningsområden för digitala vattenmärken behandlats:

Identifikationsmärkning är robusta vattenmärken som används för att lagra information, exempelvis om vilken sensor som skapat data eller information om vem som är ägaren av en bild. Mängden information som vattenmärket innehåller påverkar dels robustheten (precis som för informationsöverföring i en kanal) men även hur stor påverkan blir på den sensordata som vattenmärket är inbäddad i. En undersökning av hur vattenmärkta bilder påverkar en måldetektionsalgoritm visar att ju större och starkare vattenmärket är, desto sämre blir detektionsresultatet. Detta kommer från den distorsion som vattenmärket inför i bilden. Så länge som vattenmärket inte inför någon visuell distorsion är dock påverkan på detektionsresultatet minimalt. Därmed verkar det vara möjligt att även vattenmärka data som skall användas för automatiska signalbehandlingsalgoritmer.

Individuella vattenmärken är unika för varje användares kopia av sensordata. Om exempelvis en bild har spridit sig vidare olovligt går det att genom att undersöka dess individuella vattenmärke spåra spridningskällan. Eftersom varje auktoriserad användare då behöver ha tillgång till unik sensordata blir lagring och överföring av sådan data ett intressant problem. Vi visar att det finns flera olika sätt att effektivt överföra bilder och video så att varje mottagare får ett individuellt vattenmärke inbäddat. Detta fungerar inte bara för multicast och broadcast utan även i peer-to-peer och ad hoc-nätverk. Speciellt vår Modified Transform Watermarking-metod visar sig vara effektiv vid överföring och lagring. Den visar sig ge vattenmärken som är robusta för stillbilder som attackeras av en ensam eller flera samarbetande användare, med avsikt att radera det inbäddade märket.

Förändringsindikation kan uppnås med hjälp utav bräckliga vattenmärken som går sönder om bilden manipuleras. I ett senare skede går det att kontrollera om vattenmärket i sensordata är helt. Om det är skadat vet man att sensordatat inte längre är korrekt. Genom vår implementation av en autentiseringsalgoritm visar vi hur det går att upptäcka inom vilka områden som en stillbild har blivit förändrad. Metoden kan för övrigt anpassas för att inte reagera på vissa typer av små förändringar som JPEG-komprimering. Vi har dessutom demonstrerat att det går att delvis återskapa skadade områden via att en lågupplöst version av bilden finns inbäddad som ett vattenmärke.

Genom implementation och tester har vi påvisat möjligheten att använda vattenmärken för att bädda in metadata, användar-ID och autentiseringsinformation i sensordata och därmed höja dess sekretess, integritet och spårbarhet.

Referenser

Anderson, R., & Manifavas, C. (1997, January). Chameleon - A new kind of stream cipher. In Proceedings of the Fourth International Workshop on Fast Software Encryption FSE'97 (pp. 107–113).

Bay H., Ess A., Tuytelaars T. & Gool L. V. (2008). SURF: Speeded Up Robust Features, Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346-359.

Bennet F., & Fenelius S. (2003, March). SceneServer - a 3D software assisting developers of computer vision algorithms, FOI-rapport FOI-R--0831--SE.

Brown, I., Perkins, C., & Crowcroft, J. (1999, November). Watercasting: Distributed watermarking of multicast media. In Proceedings of the First International COST264 Workshop on Networked Group Communication NGC'99, (pp. 286–300).

Cox, I. j., Miller, M. L., Bloom, J. A., Fridrich, J., & Kalker, T. (2008). Digital Watermarking and Steganography – Second Edition. Morgan Kaufmann.

Emmanuel, S., & Kankanhalli, M. S. (2003). A digital rights management scheme for broadcast video. Multimedia Systems Journal, 8, 444–458. ACM-Springer Verlag.

Hartung F. & Kutter M. (1999, July). Multimedia watermarking techniques, In Proceedings IEEE: Special Issue on Identification and Protection of Multimedia Information, volume 87(7), pages 1079–1107.

Holliman M. & Memon N., (1997, March). Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes, IEEE Trans.Image Processing, vol. 6, pp. 432–441.

Judge, P. & Ammar, M. (2000, June). WHIM: Watermarking multicast video with a hierarchy of intermediaries. In Proceedings of the International Workshop on Network and Operation System Support for Digital Audio and Video NOSSDAV'00, 39(6), 699-712.

Kundur D. & Hatzinakos D. (1999, July). Digital Watermarking for Telltale Proofing and Authentication, Proceedings of the IEEE, vol. 87, no. 7 pp. 1167-1180.

Kundur, D. & Karthik, K. (2004, June). Video fingerprinting and encryption principles for digital rights management. Proceedings of the IEEE, 92(6), 918–932.

Lin C-Y. & Chang S-F., (2000). Semi-Fragile Watermarking for authenticating JPEG Visual Content, SPIE Security and Watermarking of Multimedia Contents II, EI '00, San Jose, CA.

Luh, W., & Kundur, D. (2005). New paradigms for effective multicasting and fingerprinting of entertainment media. In IEEE Communications Magazine, 43(6), 77–84.

Näsström F., Cirkic M., Rydell J., Skoglar P., Stenborg K-G. & Ulvklo M., (2009, March). Autonoma funktioner för intelligenta operatörsstöd, FOI-rapport FOI-R--2752--SE.

Parviainen, R., & Parnes, P. (2001, May). Large scale distributed watermarking of multicast media through encryption. In International Federation for Information Processing Communications and Multimedia Security Joint working conference IFIP TC6 and TC11.

Rivest R. L. (1992, April). The MD5 message digest algorithm, Tech. Rep., RFC 1321.

Stenborg, K-G., Herberthson, M. & Haapalahti, G. (2008, December). Identifiering och förändringsindikering av sensordata via vattenmärken, FOI-rapport FOI-R--2699--SE.

Stenborg, K-G. (2009 March). Scalabale Distribution of Watermarked Media,. Handbook of Research on Secure Multimedia Distribution, Chapter 18, 335-351, Information Science Reference, IGI Global.

Stenborg K-G. (2009 October). Individually Watermarked Information Distributed Scalable by Modified Transforms, RTO-MP-IST-087: Information Management and Exploitation, Paper 16, Stockholm.

Stenborg K-G. & Haapalahti G. (2009, December). Vattenmärkning av sensordata - Årsrapport 2009, FOI-rapport FOI-R--2943--SE.

Stenborg K-G. & Forchheimer R. (2010, October). Individual Watermarking over P2P Networks, Technology and Methodology for Security and Crisis Management TAMSEC'10.

Stenborg K-G., Herberthson M. & Forchheimer R. (2011, February). Distribution of Individually Watermarked Content in Peer-to-Peer Networks, in 4th IFIP International Conference on New Technologies, Mobility and Security NTMS'11, Paris. Kommande publikation.

Sun Q. & Chang S-F. (2005). Signature-based Media Authentication, Chapter 21 in Multimedia Security Handbook, Edited by B. Furht and D. Kirovski, CRC Press.

Viola P. & Jones M. (2001). Robust real-time object detection, Technical Report CRL 2001/01, Cambridge Research Laboratory, Compaq Computer Corporation.

Wu, T-. L., & Wu, S. F. (1997). Selected encryption and watermarking of MPEG video. In International Conference on Image Science, Systems, and Technology CISST'97.

Zhao H. V., & Liu K. J. R. (2006, January). Fingerprint Multicast in Secure Video Streaming. In IEEE Transactions on Image Processing, Vol. 15, No 1 (pp. 12-29).