CAPITOLO UNDICESIMO
DIGITAL VIDEO FORENSICS: STATUS E PROSPETTIVE

Sebastiano Battiato - Giovanni Maria Farinella
Giuseppe Messina - Giovanni Puglisi

1. INTRODUZIONE

La crescente diffusione di dispositivi di *imaging* a basso costo e la conseguente disponibilità di grosse quantità di foto e filmati digitali rende l'attività investigativa e di verifica su tali supporti sempre più frequente. Solo nel mondo virtuale della Rete stime recenti, aggiornate a metà del 2009, indicano in circa 4 miliardi il numero di foto presenti su *Flickr* (www.flickr.com) con circa 4000 upload al minuto. Su *YouTube* (www.youtube.com) si stimano circa 20 ore di upload al minuto per un totale di circa 120 milioni di video. Infine su *Facebook* (www.facebook.com) si procede al ritmo di circa 22000 upload al minuto per un totale stimato di 15 miliardi di immagini. A questi dati possiamo aggiungere inoltre l'immensa mole di informazioni multimediali per così dire "private", non necessariamente disponibili in rete nonché per esempio i dati provenienti da sistemi di video sorveglianza più o meno affidabili. L'ammontare della cosiddetta impronta digitale fa sì che per ogni essere umano sulla terra si producano al momento più di 45 Gigabyte di informazioni digitali [1]. Queste cifre ci danno l'esatta dimensione del fenomeno e devono indurre gli addetti ai lavori ad una pronta riorganizzazione delle tecniche di indagini. A tale scopo, risulta fondamentale riuscire ad individuare con esattezza modalità di analisi e di studio di tali reperti al fine di non lasciare nulla di intentato nella ricerca di fonti di prova spesso decisive. Come recentemente affermato in uno *special issue* su "*Digital Forensics*" della rivista *IEEE Signal Processing Magazine* [2] le potenzialità indotte dall'utilizzo consapevole delle informazioni contenute in un segnale digitale (audio, immagine, video, ecc.) sono notevoli a patto che non ci si dimentichi di utilizzare i mezzi e le strategie di analisi più adeguate; ciò presuppone la conoscenza tecnica dei fondamenti di base della materia per ciò che riguarda il trattamento e la codifica dei dati multimediali e che non possono essere lasciati al caso o all'improvvisazione.
Il testo di G. Reis [3] rappresenta il primo tentativo, sia pur limitato, volto a

rendere disponibile una "pratica guida" agli investigatori e ai professionisti del settore nell'ambito dell'*image and video forensics*. A tutt'oggi infatti non esiste un vero e proprio manuale di riferimento, in quanto troppe sono le lacune sia dal punto di vista teorico metodologico che dal punto di vista degli strumenti software esistenti che contraddistinguono i prodotti esistenti sul mercato. Uno dei contributi di questo lavoro consiste appunto nel fornire alcune nozioni tecnico-metodologiche per il trattamento dei video digitali in ambito forense. Verranno inoltre riportati una serie di casi esemplificativi volti a sottolinearne la potenzialità di base e i possibili scenari applicativi che si stanno affermando nella letteratura di riferimento.
Dal sito dell'FBI riportiamo la seguente definizione: "*Forensic Image analysis is the application of image science and domain expertise to interpret the content of an image or the image itself in legal matters*". In sostanza per *image/video forensics* si intende tutto ciò che riguarda l'attività di analisi delle immagini e dei video (digitali o meno) per la validazione in ambito forense. Rientrano in questo contesto le tecniche di analisi volte al miglioramento e/o al ripristino di informazioni [4], all'identificazione della sorgente di acquisizione [5],[6], all'individuazione di eventuali *forgery* o manomissioni [7],[8], ecc. In questo documento ci concentreremo sui video digitali e sulle potenziali applicazioni cercando di distinguere tra le tecniche puramente accademiche, al confine della ricerca, rispetto a ciò che invece è già disponibile e pronto per essere utilizzato in ambito forense.

2. FONDAMENTI DI ELABORAZIONE VIDEO

Una sequenza video può essere sintetizzata in due modi: o come segnale discreto che attua un campionamento temporale della scena reale (ovvero ad ogni istante la scena è "fotografata"); oppure attraverso la successione di istantanee appositamente composte. Tale sequenza è quindi costituita da una serie di *frame* cioè dalle singole immagini che compongono il video dette anche fotogrammi. Oltre che dalla necessità di contenere le dimensioni delle immagini, al fine sia di memorizzarle che di trasmetterle, la compressione di sequenze video trae origine dalla necessità di garantire la riproduzione delle sequenze con un adeguato *framerate*. Il *framerate* o *frame* (o immagini) per secondo, è l'unità di misura della frequenza di visualizzazione delle singole immagini che compongono il video. Quando la visualizzazione del filmato è *online*, rispetto alla trasmissione, come in alcune applicazioni killer quali la TV digitale o i sistemi di videoconferenza, è fondamentale mantenere un buon *framerate*. Per esempio, lo standard TV PAL richiede la riproduzione, 25 volte al secondo, di *frame* di dimensioni pari a 768 (colonne) x 576 (righe). A 24

bits per pixels (*bpp*) questo darebbe luogo, anche trascurando altri aspetti non secondari della elaborazione del segnale TV, ad un *bitrate* minimo di 768 x 576 x 25 x 24 ≈ 31 *Mbyte/sec*, assolutamente improponibile per usi comuni.
L'estensione di metodi di compressione, come quelli per le immagini fisse, ad esempio il formato JPEG, porta ad algoritmi di codifica che, sfruttando solo la ridondanza spaziale, non garantiscono qualità sufficiente soprattutto a bassi *bitrate*. Il termine *bitrate* viene solitamente utilizzato a proposito di scambi di informazioni tra computer o comunque dispositivi elettronici. Su questi dispositivi l'informazione viene memorizzata e viaggia in forma digitale, ovvero in *bit*; di conseguenza la velocità di trasmissione si misura in *bit* per secondo (e da qui il termine equivalente inglese *bitrate*).
Per poter trasmettere o memorizzare dei file video è necessario definire degli standard riguardanti sia gli algoritmi di codifica/decodifica dei flussi multimediali, sia i protocolli necessari al loro trasferimento e al loro controllo sulla rete. Gli standard cosiddetti "aperti" prevedono:

- La definizione e adozione di specifiche universalmente accettate;
- La definizione dei formati per dati multimediali e degli algoritmi utilizzati per codificarli.

In un video i singoli *frame*, che costituiscono la sequenza non sono indipendenti; in genere, la scena cambia lentamente e in parte, ovvero esiste una coerenza temporale tra fotogrammi vicini, e una ridondanza nel tempo (tra *frame*). Per questo motivo si usa tracciare il moto degli oggetti nella scena, al fine di ridurre le ridondanze *inter-frame*.
Un *CoDec* video (*Co-Dec = enCoder/Decoder*) è un software composto da due parti: un *enCoder* che comprime la sequenza di immagini archiviandola in un file e un *Decoder* necessario per decomprimere la sequenza e poterla nuovamente visualizzare. A loro volta le tecniche di compressione video possono essere suddivise in: tecniche *lossless*, dove la compressione è un processo perfettamente reversibile che avviene senza perdita di informazione; e tecniche *lossy* dove la compressione non è reversibile, in questo caso i video compressi e decompressi non sono più perfettamente identici in quanto al momento della compressione sono state volutamente eliminate alcune informazioni ritenute "sacrificabili". Per comprimere il video si utilizzano tecniche che sfruttano alcune caratteristiche intrinseche del video stesso, in combinazione con le caratteristiche del sistema visivo umano. In particolare è possibile comprimere un segnale video attaccando la ridondanza spaziale, la ridondanza temporale e sfruttando le caratteristiche del sistema visivo umano.
Rimuovendo la ridondanza statistica (ripetitività) contenuta in un video e mantenendo solo le informazioni effettivamente utili, si cerca una rappresentazione "meno correlata" delle immagini, eliminandone quindi le "ripetizioni". Si può dimostrare che *pixels* adiacenti, vicini, all'interno di una

stessa immagine, presentano caratteristiche molto simili per quel che riguarda il colore e la luminosità. La "*Codifica Intra-Frames*" si occupa di rimuovere questa ripetitività altresì detta "Ridondanza Spaziale" all'interno dello stesso fotogramma. Come accennato precedentemente esiste inoltre una netta correlazione anche tra i *pixels* di fotogrammi adiacenti. Un fotogramma ed i due vicini (il successivo ed il precedente) spesso risultano pressoché identici (fanno eccezione le situazioni in cui si hanno cambi di scena). Questa "Ridondanza Temporale" tra fotogrammi vicini, che ne sfrutta le loro minime differenze, viene trattata dalla "*Codifica Inter-Frames*". Infine, sfruttando alcune peculiarità del sistema visivo umano, ovvero la scarsa sensibilità dell'occhio alle alte frequenze video soprattutto in presenza immagini in movimento è possibile "tagliare", buttar via, alcune informazioni (le alte frequenze cioè i cambiamenti repentini o i dettagli molto fini) di un'immagine senza introdurre artefatti visibili. Il sistema visivo umano non è, infatti, in grado di percepire le variazioni nei dettagli di figure molto frastagliate. E' ad esempio molto difficile rendersi conto di una perdita di dettaglio nelle fronde di alcuni alberi in movimento mentre è quasi immediato notare anche la più piccola variazione di colore o luminosità nell'azzurro di un cielo limpido e sereno sullo sfondo di un video.

Il primo passo nella rimozione della "Ridondanza Spaziale" consiste nell'attuare una compressione statica dei dati seguendo una codifica simile a quella di tipo JPEG: conversione del *frame* dallo spazio di colore RGB a quello YUV; suddivisione del *frame* in macroblocchi di dimensioni 16x16 *pixels;* applicazione della trasformata Discreta del Coseno (DCT) su ognuno dei blocchi; quantizzazione dei coefficienti DCT; scansione a Zig-Zag per l'allineamento dei coefficienti quantizzati. Considerando quindi la suddivisione del *frame* in blocchi non sovrapposti si esplicita l'idea di rendere uniforme il moto di pixel vicini (all'interno di ogni blocco). Ad ognuno dei blocchi così ottenuti è assegnato un vettore che ne identificherà il moto *intra-frame*. Si noti che la ricerca per il blocco di riferimento viene generalmente effettuata in un intorno limitato del blocco di partenza in quanto se la ricerca fosse effettuata su tutto il fotogramma, i tempi per la codifica potrebbero allungarsi in modo non accettabile. L'efficienza del metodo è assicurata dal principio della "Ridondanza Spaziale": la probabilità di trovare un blocco simile man mano che ci si allontana dal blocco di partenza, diminuisce in modo esponenziale con l'aumentare della distanza.

Per quanto riguarda la riduzione di "Ridondanza Temporale" si attua una compressione dinamica dei dati. A tal fine occorre effettuare da prima una stima del moto (*Motion Estimation* - ME) dei blocchi tra frame adiacenti e quindi procedere con la compensazione di questo moto (*Motion Compensation* - MC). Anche in questo caso il frame è suddiviso in macroblocchi di dimensioni 16x16 *pixels*, si applica la trasformata Discreta del Coseno (DCT) su ognuno

dei blocchi; si quantizzano i coefficienti DCT e infine si effettua una scansione a Zig-Zag per l'allineamento dei coefficienti quantizzati.
L'encoder individua tra i fotogrammi adiacenti (nel solo fotogramma precedente, o nel precedente e nel successivo a seconda dei casi) il blocco più simile (se non uguale) e associa al blocco su cui è stata effettuata l'analisi, un vettore di moto, cioè una coppia di numeri (X,Y) che individuano sul piano ipotetico rappresentato dal fotogramma, il vettore di spostamento, che indica verso e entità dello spostamento del blocco passando dal fotogramma 1 al fotogramma 2. Viene quindi creato il blocco differenza, in pratica viene sostituito il blocco originale su cui è stata effettuata la ricerca, con il risultato che si ottiene sottraendo dal blocco molto simile trovato del fotogramma 1 (*frame* di riferimento) il blocco in questione nel fotogramma 2.
Se tutto ha funzionato a dovere, ovvero l'encoder non ha commesso errori nella ricerca del blocco di riferimento (nel fotogramma 1), si ottiene un netto vantaggio consistente nel fatto che il nuovo blocco "differenza" sarà costituito da un numero decisamente inferiore di dati. Saranno infatti presenti molti zeri, l'entropia sarà minore, la codifica più efficiente. Le uniche informazioni aggiuntive da considerare saranno quelle relative al vettore di moto (una coppia di numeri).

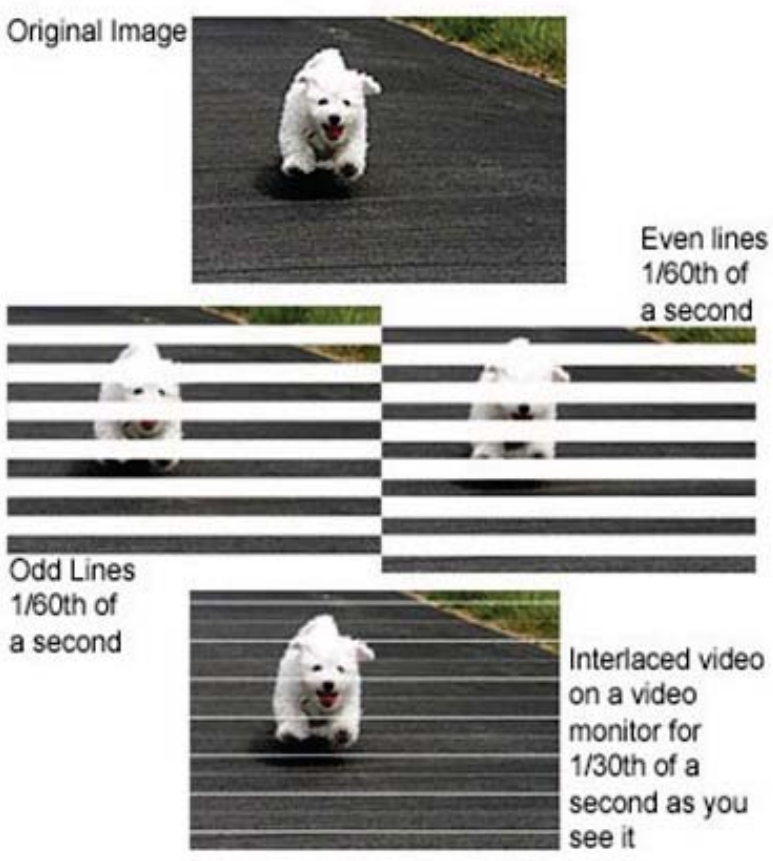


Figura 1 - Esempio di immagini interlacciate.

2.1. Interlacciamento

L'interlacciamento è un sistema di scansione di immagini video che prevede la divisione delle linee di scansione in due parti, dette campi, semiquadri oppure *field*, suddivisi in linee pari e dispari. Questa tecnica permette una qualità di trasmissione migliore senza bisogno di aumentare la larghezza di banda. Un televisore in standard PAL, per esempio, visualizza 50 *field* al secondo (25 pari e 25 dispari) poiché la sua trasmissione avviene a 50 Hz (*Hertz*). Un *frame* completo, quindi, viene tracciato 25 volte al secondo. I monitor a tubo catodico per uso televisivo sono sempre stati interlacciati fino alla fine degli anni 70 quando l'avvento dei personal computer e delle relative risoluzioni ha comportato la reintroduzione di CRT a scansione progressiva.
L'interlacciamento è anche previsto dai formati per l'alta definizione (HD), in particolare dal 1080i. I moderni televisori al plasma e LCD sono tutti di natura progressiva, mentre monitor CRT interlacciati si utilizzano in ambito professionale per l'equipaggiamento di sistemi di produzione e controllo. L'interlacciamento è una tecnica particolarmente utile per ridurre la banda del segnale di un fattore pari a due, indipendentemente dal numero di linee e dalla frequenza di visualizzazione. Con una data larghezza di banda è possibile riprodurre un segnale video interlacciato con una frequenza di visualizzazione doppia. Una frequenza di visualizzazione elevata riduce lo sfarfallio nei monitor CRT e migliora la visualizzazione dei movimenti, aumentando la risoluzione temporale del sistema di scansione. Il sistema visivo umano calcola la media tra i fotogrammi rapidamente visualizzati in un video, e quindi gli artefatti generati dall'interlacciamento non sono percepibili se trasmessi alla velocità corretta.
A parità di banda e di frequenza di scansione, il video interlacciato permette di ottenere una risoluzione spaziale più alta rispetto a quella della scansione progressiva. Per esempio, un segnale in alta definizione 1080i50, interlacciato con risoluzione 1920x1080 e frequenza di 50 Hz, occupa una banda simile a un segnale 720p50, a scansione progressiva con risoluzione di 1280x720 e frequenza 50 Hz. Il primo segnale però ha circa il 50% in più di risoluzione spaziale.
Il video interlacciato è progettato per essere acquisito, registrato, trasmesso e visualizzato sempre restando interlacciato. Il limite principale della tecnica dell'interlacciamento è la creazione di artefatti visibili durante i movimenti rapidi, in particolare quando dei pezzi della scena (soggetti o oggetti) si muovono abbastanza velocemente da essere in due posizioni diverse nei due *field* di uno stesso fotogramma.
Gli artefatti sono facilmente visibili durante la riproduzione di immagini fisse o a velocità inferiori a quella nominale [9].

2.2. Lo standard MPEG

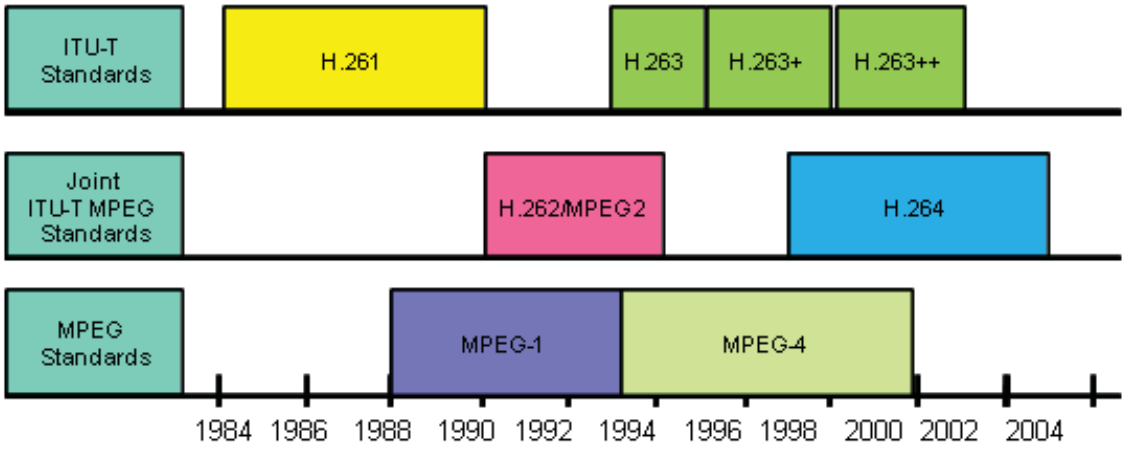


Figura 2 - Scala temporale di sviluppo dei diversi standard video.

Nel 1988 l'organismo internazionale di standardizzazione ISO-IEC (*International Organization for Standardization and International Electrotechnical Commission*) si è assunto il compito di sviluppare uno standard per la compressione e la rappresentazione del video digitale e dell'audio ad esso associato che fosse adatto alla memorizzazione su dispositivi di memoria di massa (dischi ottici, DAT) e alla trasmissione su canale di telecomunicazione (ISDN, LAN, TV). Il *Moving Picture Expert Group* (MPEG) è il comitato internazionale nato in seno all'ISO per raggiungere tale obiettivo. Formalmente MPEG è il gruppo di lavoro 11 del sub-comitato 29 del *Joint Technical Group* 1 dell'ISO-IEC.

2.3. Il formato MPEG-1

Questo formato è nato per rispondere all'esigenza di memorizzare filmati su compact disc. E' stato definito come standard nel 1992 ed utilizza un *bitrate* costante di 1,15 Mbit/sec per il video e dai 384 ai 198 Kbit/sec per l'audio. Da un punto di vista qualitativo, l'obiettivo prefissato era il raggiungimento della qualità VHS. In pratica, per ottenerla, il video è codificato a 352x288 pixel (288 linee orizzontali da 352 punti ciascuna) per quanto riguarda la luminosità, mentre per quanto riguarda il colore l'immagine è ulteriormente divisa per due ed è pertanto codificata a 176x144. Per ottenere il video con la qualità SIF(352x240), il *codec* MPEG-1 effettua una serie di operazioni di compressione delle immagini che sfruttano non solo l'algoritmo DCT, ma anche le differenze tra un fotogramma e l'altro. Anziché memorizzare tutti i fotogrammi per intero, se ne memorizzano soltanto alcuni come tali (ad intervalli prefissati e regolari), e tra di essi ci si limita a memorizzare una serie di *frames* incompleti nei quali vengono "scritte" solo le informazioni che subiscono delle variazioni rispetto

alle immagini precedenti. MPEG-1 gestisce solo *frame* (*progressive scan*) e non è stato progettato per la gestione di sequenze di tipo interlacciate.

2.4. Frames I/P/B

Gli standard MPEG prevedono la classificazione dei *frame* in tre tipi: "I", "B" e "P". Il *frame* "I" è un frame video completamente indipendente. Il *frame* "P"(*predictive frame*) si basa su un precedente *frame* "I". Il *frame* "B" (*Bidirectional frame*) è costituito da informazioni ricavate sia da *frame* "I"che da *frame* "P" (anche successivi) attraverso interpolazione.

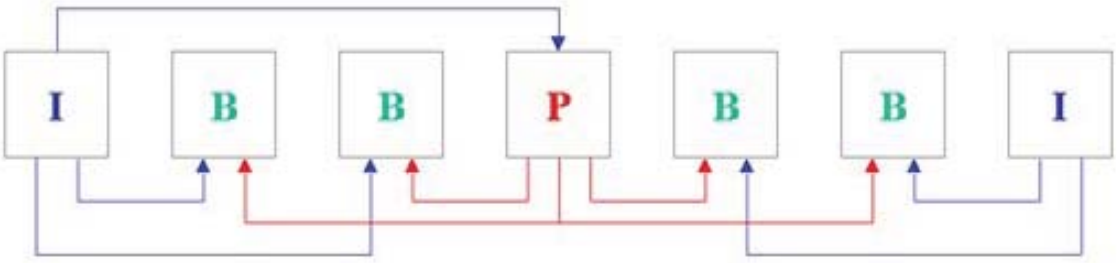


Figura 3 - Disposizione tipica di una sequenza I/P/B.

I fotogrammi di tipo "I", chiamati anche *Intra-Frames* o *Key-Frames* (fotogrammi chiave), sono fotogrammi che vengono codificati utilizzando i principi della "Ridondanza Spaziale" e non contengono nessun riferimento o informazione sui fotogrammi adiacenti. In pratica sono compressi alla stregua di un'immagine singola, allo stesso modo di quando un'immagine viene salvata in formato JPEG. Nessun tipo di compressione temporale (ovvero compressione che tiene conto anche dei fotogrammi successivi e/o precedenti) viene applicata a questi fotogrammi.
In genere i fotogrammi chiave vengono inseriti dal *codec* ogni qualvolta vi sia un repentino cambiamento tra due immagini successive. Se inoltre viene specificato un intervallo massimo tra un fotogramma chiave ed il successivo, il *codec* dovrà necessariamente inserire un fotogramma chiave anche se non strettamente necessario.
Il fotogramma di tipo "P" (*Predicted frames*) è codificato utilizzando informazioni acquisite in base al fotogramma che lo precede, sia questo di tipo "I" o di tipo "P". Ogni macroblocco di 16x16 pixels di un *P-Frame* può essere codificato in modo indipendente (come nel caso dell'*I-Frame*) oppure può essere compensato, cioè bilanciato utilizzando informazioni del fotogramma precedente. Utilizzando le somiglianze tra fotogrammi successivi i fotogrammi "P" risultano essere più piccoli dei corrispondenti *I-Frames*. Un fotogramma di tipo "P" contiene le informazioni della posizione (X',Y') nel fotogramma

corrente in cui si è spostato un blocco che aveva coordinate (X,Y) in quello precedente (*Motion Estimation/ Compensation*). Lo svantaggio dell'utilizzo di questo tipo di fotogrammi si ha in fase di decodifica; è infatti necessario "ricostruire" ciascun fotogramma P prima di poterlo visualizzare, e per far questo si deve sempre partire dal fotogramma P seguente all'ultimo fotogramma chiave.
Per i fotogrammi di tipo "B" la ricerca del moto (*Motion Estimation/ Compensation*) è effettuata non solo sul fotogramma precedente (come nel caso di *P-Frames*) ma anche sul fotogramma successivo. La codifica ed anche la decodifica risultano quindi decisamente più complesse. Sostanzialmente i fotogrammi "B" sono di tipo "Bidirezionale", nel senso che fanno riferimento sia a ciò che li precede, sia a quello che segue. Inserire in un fotogramma informazioni che si riferiscono ad un fotogramma successivo è possibile solo alterando l'ordine in cui i fotogrammi vengono archiviati all'interno del file video compresso.

2.5. Il formato MPEG-2

Questo standard è stato sviluppato partendo dall'MPEG-1 ed esiste dal 1994. L'obiettivo dell'MPEG-2 era quello di creare un formato flessibile ed adatto a varie applicazioni, in grado anche di codificare in digitale le immagini con una qualità equivalente a quella analogica definita come broadcast (corrispondente alla qualità delle trasmissioni televisive), e l'audio con quella cinematografica, utilizzando flussi di dati fino a 60 Mbit/sec.
La caratteristica principale dell'MPEG-2 è la sua scalabilità, ovvero la possibilità di creare soluzioni di codifica e decodifica più o meno complesse in base al tipo di prodotto da realizzare, aggiungendo poi altre caratteristiche quali la possibilità di trasmettere il flusso multimediale su reti a larga banda, assicurando una buona robustezza nei confronti degli errori della rete, il trasporto parallelo di molteplici canali audio, le funzioni di protezione e di controllo di accesso al flusso, solo per citarne i principali.
Per consentire all'industria di procedere gradualmente con l'implementazione dello standard, il comitato di lavoro dell'MPEG ha definito una serie di livelli e di profili in base ai quali ogni soluzione tecnica può essere sviluppata e verificata. Non tutte le combinazioni portano ad un sottoinsieme di specifiche valide, per questo i cinque profili ed i quattro livelli si combinano solo in 11 soluzioni (e non già venti).

Profilo	Livello	Pixel Orizzontali	Pixel Verticali	Frame Rate Max	Bitrate Max (Mbit/sec)
Simple	main	720	576	30	15
Main	low	352	288	30	4
Main	main	720	576	30	15
Main	high 1440	1440	1152	60	60
Main	high	1920	1152	60	80
SNR Scalable	low	352	288	30	3(4)
SNR Scalable	main	720	576	30	10(15)
Spatially Scal.	high 1440	720 (1440)	576 (1152)	30 (60)	15(40 o 60)
High	main	352 (720)	288 (576)	30 (30)	4 (15 o 20)
High	high 1440	720 (1440)	576 (1152)	30 (60)	20 (60 o 80)
High	high	960 (1920)	576 (1152)	30 (60)	25 (80 o 100)

Tabella 1 – Profili MPEG-2.

I profili sono: *Simple (SP), Main (MP), SNR, Spatial Scalable, High*. Mentre i livelli sono *Low (LL), Main (ML), High 1440 (H-14), High (HL)*. MPEG-2 introduce i concetti di *frame picture* e di *field picture*, associando ad essi i metodi di codifica basati sulla predizione sul *frame* e predizione sul *field*. Questo formato è largamente utilizzato nei film in formato DVD.

2.6. Il formato MPEG-4

L' MPEG-4 usa fondamentalmente lo stesso algoritmo di compressione di MPEG-1 e MPEG-2, ma in modo molto più efficiente. La differenza sostanziale è che il sistema riesce a distinguere i vari livelli di un immagine, lo sfondo e i primi piani. Se lo sfondo rimane uguale nel fotogrammi successivi non verranno memorizzati, risparmiando così prezioso spazio. Inoltre è possibile elaborare queste immagini più semplicemente, estrapolando gli attori o gli oggetti dallo sfondo con grande facilità.

2.7. Il formato H.264

L'MPEG-4 Part 10, designazione formale ISO/IEC 14496-10, comunemente chiamato MPEG-4 AVC, anche abbreviato AVC (acronimo di *Advanced Video Coding*) designazione ITU-T H.264 è ciò che nel gergo tecnico comune si intende per codifica H.264. In altri termini è una versione avanzata del formato

MPEG-4 ottenuto dal lavoro congiunto del *ITU-T Video Coding Experts Group* (VCEG) ed del *ISO Moving Picture Experts Group.* A parità di *bitrate*, la qualità percepita e la risoluzione raddoppiano rispetto al formato MPEG2. Per ottenere questo risultato è stata però incrementata la complessità di codifica e decodifica dei video. Questo formato è progettato per i supporti HD-DVD e *Blue-Ray* e ne è prevista l'adozione anche per la televisione HDTV. L'H.264 si rivolge ad un grande numero di applicazioni, inclusa la trasmissione di contenuti video a basso *bitrate* attraverso le reti *wireless*, il *video streaming* attraverso Internet, la distribuzione di contenuti in qualità DVD via *broadband* o il cinema digitale.
Rispetto allo standard MPEG-2 è stato introdotto un miglioramento sfruttando tante piccole ottimizzazioni locali a tecniche esistenti: ovvero ottimizzazione della trasformata DCT, delle tabelle di quantizzazione, della codifica entropica, ecc. E' inoltre stato aggiunto un filtro per la riduzione dell'effetto *blocking.*
Nel settore della videosorveglianza, è altamente probabile che lo standard di compressione H.264 venga rapidamente adottato per applicazioni che richiedono risoluzioni e velocità di trasmissione elevate, ad esempio per la sorveglianza di autostrade, aeroporti e casinò, dove l'uso di 30/25 fotogrammi (NTSC/PAL) al secondo rappresenta la norma. Questi ultimi sono infatti gli ambiti in cui la riduzione della larghezza di banda e dello spazio di memorizzazione necessario può offrire i vantaggi più significativi. Lo standard H.264 è destinato probabilmente anche ad accelerare la diffusione delle telecamere di rete con risoluzione megapixel poiché questa tecnologia di compressione ultra-efficiente è in grado di ridurre le dimensioni dei file grandi e la velocità di trasmissione in bit senza compromettere la qualità delle immagini. Il nuovo standard presenta tuttavia anche degli svantaggi. Benché offra vantaggi significativi in termini di larghezza di banda e spazio di memorizzazione, questo standard richiede l'implementazione di telecamere di rete e stazioni di monitoraggio ad alte prestazioni.

3. I DETTAGLI TECNICI UTILI ALLE INDAGINI

L'estrazione di informazioni da un video digitale (per esempio un numero di targa, un volto, un tatuaggio, ecc.), a prescindere dalla sorgente di acquisizione (video camera, telefonino, sistema di video sorveglianza, ecc.) è legata principalmente ai seguenti parametri di codifica: risoluzione video e fattore di compressione (o *bitrate*). Come accennato nel paragrafo precedente, i diversi standard di codifica e le differenti tipologie di dispositivi atti alla registrazione (compresi i famigerati DVR – *Digital Video Recorder* dei sistemi odierni di videosorveglianza) non sono affatto orientati (purtroppo) all'analisi forense. Il risparmio di banda o di memoria fisica ha, infatti, imposto delle situazioni de-

facto, per cui pur avendo la possibilità di acquisire e quindi registrare immagini/video di ottima qualità da cui estrapolare utili informazioni investigativi, ci si ritrova a che fare con dei *settings* (spesso preimpostati) tali da inficiare una qualsivoglia possibilità reale di indagine.

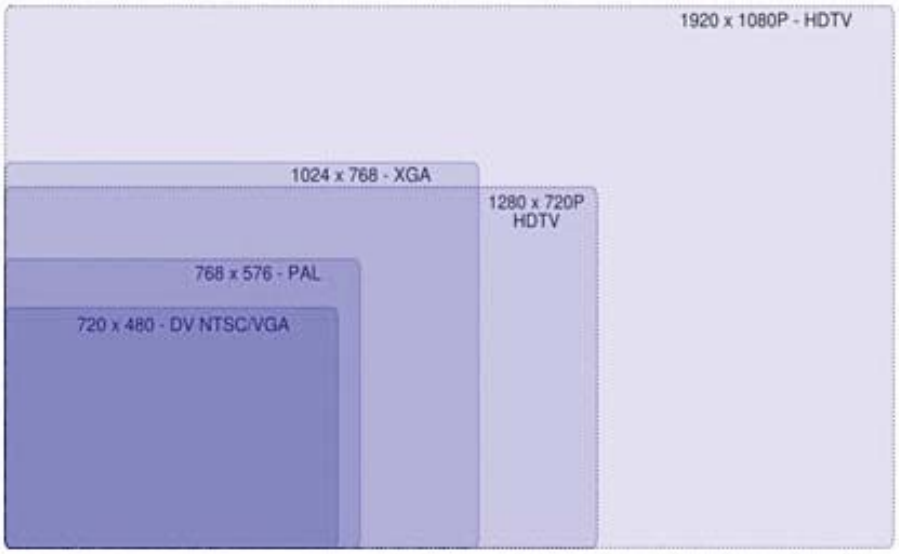


Figura 4 - Esempio di risoluzioni video attualmente esistenti.

In Figura 4 sono messe a confronto le diverse risoluzioni video oggi esistenti ed utilizzate dagli standard televisivi. Risulta ovvio come al crescere della risoluzione, le corrispondenti sequenze aumenteranno il livello di dettaglio consentendo un'accurata ricostruzione di particolari spesso decisivi.

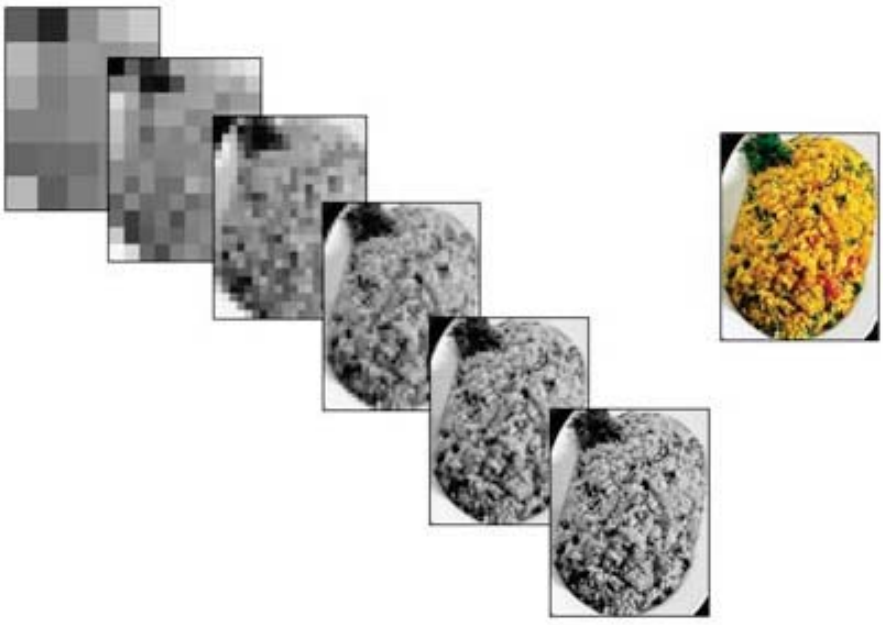


Figura 5 - Incremento dei dettagli al crescere della risoluzione di acquisizione.
Da questo punto di vista i dispositivi di acquisizione video più comuni, compresi i telefonini di fascia medio-alta, non vanno oltre una risoluzione di 2 o 3 megapixel che spesso risultano essere insufficienti nei casi reali per riuscire a ricostruire le informazioni mancanti (numeri di targa, ecc).
Il secondo fattore determinante è il *bitrate* di codifica che impatta direttamente sulla dimensione del file video finale ma che nei casi limite, introduce degli artefatti che a prescindere dalla risoluzione, possono definitivamente eliminare dalla scena alcuni dettagli importanti. Il tipico artefatto da compressione è il cosiddetto effetto *blocking*.

Figura 6 - Effetto di Blocking presente nelle immagini (o fotogrammi) eccessivamente compresse.

Premesso quanto sopra, in fase di analisi è possibile applicare delle tecniche di miglioramento o di restauro in grado di recuperare informazioni spesso presenti ma non visibili. Sono possibili diverse tipologie di elaborazioni [10]: nel dominio dello spazio cioè dei pixel, nel dominio della frequenza, nel dominio del tempo (es. informazioni provenienti da più fotogrammi). Per ogni tipo di problema possono essere applicate diverse tecniche con differenti prestazioni, complessità e costo computazionale.

4. APPLICAZIONI AVANZATE

Diverse sono le potenzialità delle tecniche oggi esistenti in grado di supportare le esigenze investigative. In particolare oltre al classico miglioramento di qualità, volto a scovare l'eventuale presenza di oggetti, targhe, individui e alla possibilità di inferire delle misure, segnaliamo a seguire alcune applicazioni avanzate che hanno suscitato l'interesse della comunità scientifica di riferimento.

4.1. Analisi automatica dei contenuti di video

Uno dei settori emergenti affronta il problema dell'estrazione di informazioni da archivi video di grandi dimensioni, al fine di individuare, in maniera automatica, situazioni di interesse. In un tale contesto l'efficienza e la velocità di analisi dei video sono di fondamentale importanza. Infatti, in questo caso, i vincoli di tempo non sono delimitati dal *framerate* ma dalla grande quantità di dati da analizzare.
In materia di analisi forense risulta di particolare interesse lo sviluppo di strumenti automatici per la analisi comportamentale. Questo tipo di strumento semplificherebbe in maniera significativa le indagini riducendone il tempo speso nella ricerca di eventuali segmenti di particolare interesse all'interno di lunghe sequenze video. Tra i diversi comportamenti sospetti, l'analisi della traiettoria di una persona è di rilevante importanza e può essere estratta automaticamente dai sistemi di video sorveglianza convenzionali. L'eccessiva numerosità di questo tipo di dati potrebbe aumentare rapidamente in scenari affollati, diventando de-facto impossibile da analizzare senza l'apporto di strumenti automatici [11].

4.2. Video superresolution - incremento della risoluzione

La risoluzione dell'immagine dipende soprattutto dal numero di pixel del sensore di acquisizione e dall'ottica adottata. Per aumentare la risoluzione si può agire sul sistema di acquisizione (ad esempio riducendo la dimensione dei pixel e quindi aumentandone la densità per millimetri quadri) e sulla tecnologia adottata. Seguendo tale approccio si presentano però numerosi limiti ed inconvenienti, oltre ai problemi economici e tecnologici. Da ciò nasce la necessità di poter disporre di tecniche alternative di *image processing* per l'aumento della risoluzione senza modificare e migliorare il sistema di acquisizione.
La super risoluzione non esprime semplicemente il concetto di incremento delle dimensioni spaziali di un'immagine ma ha che fare con il cosiddetto potere risolutivo, legato al incremento del dettaglio piuttosto che alle dimensioni stesse. In generale la super risoluzione è utile ogni qual volta si disponga di un numero

elevato di immagini a bassa risoluzione della stessa scena (per esempio facenti parte di una sequenza video) e si desideri un'immagine a più alta risoluzione, oppure quando non è possibile per ragioni tecniche o economiche cambiare il sistema di acquisizione.
Ci sono molti algoritmi in letteratura, ma fanno spesso troppe assunzioni, non sempre verificate nei casi reali, che nella pratica impediscono di ottenere i risultati desiderati [12]. Nel processo di acquisizione l'immagine reale è stata sottoposta a sottocampionamento e sfocatura. Si utilizzano le informazioni provenienti da diversi fotogrammi per ricostruire le informazioni perse nel processo di acquisizione. Il processo di super risoluzione avviene in due passi:

- Registrazione delle immagini: si utilizza una tecnica di stima del moto tra *frame* (con accuratezza superiore alla dimensione del pixel, detta sub-pixel).
- Ricostruzione dei dati mancanti mediante fusione delle informazioni a bassa risoluzione. Di solito si utilizzano degli approcci iterativi che tendono a minimizzare errori di approssimazione utili a stimare la qualità finale dell'immagine.

Nella figura seguente è illustrata la differenza, in termine di risoluzione, tra un tipico algoritmo di *zooming* e la stessa scena ottenuta mediante la tecnica di super risoluzione.

Figura 7 - Immagine originale.

a) Zooming b) Super-Risoluzione

Figura 8 – Incremento della risoluzione

4.3. Near Duplicate Video Identification - Identificazione di copie manipolate.

Rintracciare i produttori e distributori di materiale video illecito, in particolare per ciò che riguarda la pedopornografia, è oggi uno degli obiettivi primari da parte delle agenzie investigative di tutto il mondo. La tipica operazione di sequestro di possibile materiale illegale prevede il sequestro dei relativi dispositivi di archiviazione del sospettato. Viene quindi fatta una copia di ciascun dispositivo per evitare l'eventuale manomissione dei dati originali. Infine si procede all'analisi dei dati per individuare i video presenti e quindi passare alla conseguente identificazione.
I metodi di indagine utilizzati per questo processo, tuttavia, sono ancora molto grossolani. Le informazioni sui file video sono in genere inserite in un sistema di analisi forense (*Encase by Guidance* sembra essere l'applicazione utilizzata più comunemente). L'investigatore deve quindi aprire manualmente ogni file e controllarne il contenuto. Si noti che l'intero file deve essere sottoposto a scansione, poiché spesso il materiale illegale è inserito all'interno di sequenze legali al fine di evitarne l'identificazione. Naturalmente capita di individuare gli stessi video più e più volte, ma devono essere controllati in ogni caso, poiché spesso i nomi dei file e i contenuti sono regolarmente riorganizzati e modificati. Al fine di ridurre il carico di lavoro è stato introdotto in taluni casi l'uso del cosidetto *checksum MD5* al loro processo di indagine.
Il *checksum MD5* [13] è un valore *hash* che viene calcolato in questo caso sui dati video ed è utilizzato al fine di inserire una firma e generare una raccolta di *checksum MD5* noti sia per il materiale legale che per quello illegale. Se il *checksum* individuato corrisponde al *checksum* di un video legale, non vengono presi ulteriori provvedimenti. In caso contrario, il video deve essere acquisito. Purtroppo, però, la corrispondenza effettiva dei *checksum* è incline al fallimento. Ad esempio, modificando anche un singolo bit di un file, il *checksum* cambia totalmente. Inoltre, un video illegale può essere manipolato in modo tale che il suo *checksum* corrisponda a quello di un video legale ben noto, pur mantenendo i contenuti illegali, provocando l'esclusione del materiale dalle indagini [14]. A scopo esemplificativo si supponga di esaminare video composti da un solo *frame*[1] e che la banca dati a disposizione degli investigatori contenga il video *A* mostrato in Figura 9. Si supponga anche che durante le indagini vengano acquisiti cinque diversi video come mostrato in Figura 10. E' chiaro che, se l'analisi visuale fosse condotta da un investigatore, questi assocerebbe i video a), b), c), d) in Figura 10 al video *A* mostrato in Figura 9, mentre il video e) sarebbe considerato un video avente contenuto diverso da quello presente nel video *A*. Sebbene i video in a), b), c), d) in Figura 10 non sono identici al video

[1] I ragionamenti seguenti possono essere estesi a video reali con più di un frame.

A mostrato in Figura 9, questi risultano avere contenuti visuali molto simili a quelli del video *A* in Figura 9. In effetti, i video a) b) e c) riportati in Figura 10 sono ottenuti manipolando il video *A* mediante delle trasformazioni di scala, rotazione, cropping, mentre il video d) in Figura 10 è ottenuto inquadrando la stessa scena presente nel Video *A*, ma da un punto di vista differente. Il video *B* riportato in Figura 10 è invece un video dai contenuti differenti rispetto al video *A*. Ne segue che l'ispezione visuale dei video da parte di un investigatore porterebbe ad asserire che i video a), b), c), d) riportati in Figura 10 sono "duplicati" (o come si dice nel gergo informatico *Near Duplicate*) del video *A*, mentre il video e) ha un contenuto diverso da quello presente nel video *A*. Nonostante il caso in questione sia semplice da affontare per un umano, se ci si fosse affidati ad una tecnica di analisi automatica condotta mediante il *checksum MD5*, il risultato finale sarebbe stato errato in quanto il *checksum MD5* dei video in Figura 10 risulta essere diverso da quello del video *A* rappresentato in Figura 9. In Tabella 2 sono riportati i valori di *checksum MD5* dei video in questione. Una scansione automatica mediante *checksum MD5* non individuerebbe alcuna evidenza utile ad asserire che un "duplicato" del video *A* è presente tra i video acquisiti durante le indagini.

Video *A*

Figura 9 – Esempio di video d'archivio

a) Video *A* a cui è stata applicata una operazione di cambio di scala

b) Video *A* a cui è stata applicata una operazione di rotazione

c) Video *A* a cui è stata applicata una operazione di cropping

d) Video *A* in cui la scena inquadrata è stata ripresa da un differente punto di vista

e) Video *B*

Figura 10 – Esempi di video acquisiti durante ispezione.

Video	MD5
Video *A*	0134683B830447D506EC016D475B3A60
Video *B*	D7B0E8FA61928F2B15799829E7EF6CB0
Video *A* a cui è stata applicata una operazione di cambio di scala	1133CC0DA44A6544336C755393DAEC72
Video *A* a cui è stata applicata una operazione di rotazione	769D7B77C7250D26FD81893F2A7FA63C
Video *A* a cui è stata applicata una operazione di cropping	D72919114E643CACB233CC334A03D5BC
Video *A* in cui la scena inquadrata è stata ripresa da un differente punto di vista	952328C4972191F1901BDE25CB3034CD

Tabella 2 - Valori di *checksum MD5* dei video in Figura 9 e Figura 10. I valori sono stati calcolati utilizzando il software FastSum reperibile all'indirizzo web http://www.fastsum.com/.

In letteratura sono presenti approcci che permettono di calcolare un *hash* percettivo delle immagini e dei video [14],[15]. La rappresentazione delle immagini mediante *hash* percettivo si basa su caratteristiche che sono calcolate tenendo conto del contenuto visuale e semantico delle immagini e dei video. Le caratteristiche calcolate sui video digitali sono solitamente invarianti sia a trasformazioni geometriche (es. traslazione, rotazione, scalatura, ecc.) che fotometriche (es., luminosità, ecc.) delle immagini e dei video. Il processo di rappresentazione delle immagini e/o video mediante *hash* percettivo prevede solitamente la costruzione di un "vocabolario" di caratteristiche (dette *visual words*) a partire da immagini e video di esempio presenti in una banca dati d'archivio. Ogni immagine/video è poi rappresentato mediante un istogramma normalizzato che cattura le frequenze relative di ciascuna parola visuale del vocabolario nell'immagine/video in esame. Al fine di stabilire se una immagine/video in esame è simile ad uno presente in banca dati, si utilizzano delle misure di distanza opportune che permettono di organizzare i risultati lungo un asse orizzontale. Solitamente le immagini/video così analizzate vengono mostrati all'investigatore mediante una apposita interfaccia grafica da cui si può stabilire una soglia sulla distanza utile a stabilire se le immagini e/o i video acquisiti sono *Near Duplicate* di una immagine e/o video presente in archivio.

Ritornando al caso esemplificativo del riconoscimento dei video in Figura 10 rispetto alla banca dati in Figura 9, riportiamo di seguito nelle Figure 11-15 le rappresentazioni dei video mediante *hash* percettivo. In particolare nelle Figure 11-15 è rappresentato l'istogramma normalizzato del Video *A* a confronto con gli istogrammi normalizzati degli altri video. In Tabella 3 sono invece riportate le distanze tra la rappresentazione del Video *A* e quella degli altri video considerati come esempio.
Le misure di distanza riportate in Tabella 3 permettono di organizzare i video in esame lungo un asse orizzontale come riportato in Figura 16. Se ad esempio l'investigatore stabilisse che la soglia massima per cui considerare un video come *Near Duplicate* è pari 0,4 (in una scala da 0 a 1), allora solo i video a), b), c) e d) presenti in Figura 10 verrebbero correttamente associati al Video *A*, mentre il video e) sarebbe scartato automaticamente dal software.

Le metodologie avanzate per il calcolo di *hash* percettivo sono già in fase di sperimentazione avanzata nell'ambito dell'*image forensics*. Queste tecniche permettono di analizzare in maniera automatica oppure semi-automatica grosse quantità di dati in maniera efficiente ed effettiva e saranno sempre più utilizzate come supporto alle indagini.
I primi prodotti commerciali che implementano a vario titolo tecniche e concetti di analisi dei video digitali analoghi a quanto finora rappresentato sono i seguenti:

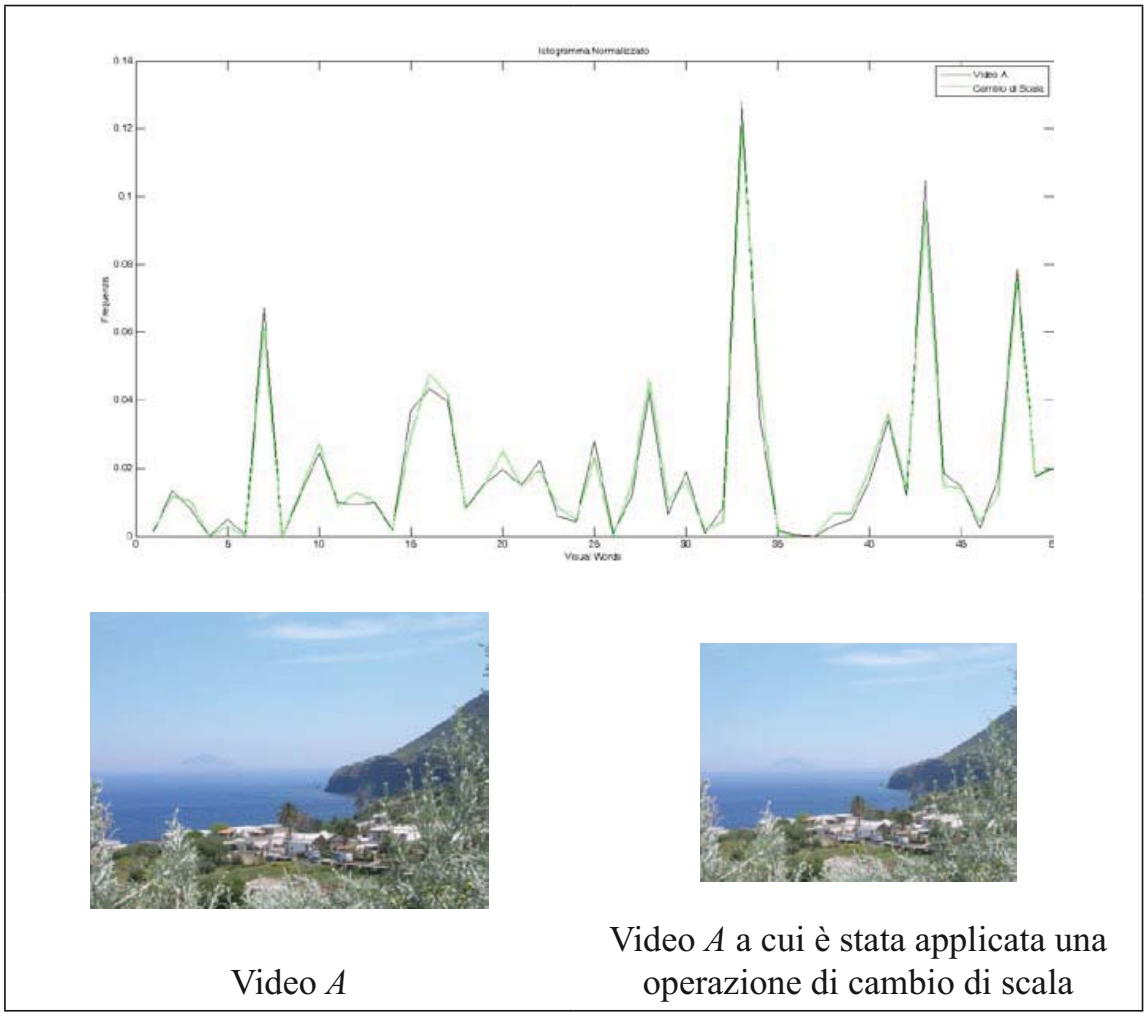
- *Videntifier* (www.eff2.net)
- *PhotoDna* (www.microsoftphotodna.com)
- *VideoGenome* (v-nome.org/)

Presso il nostro laboratorio, stiamo attualmente svolgendo delle valutazioni sul campo di alcuni dei software di cui sopra. Una particolare menzione va a *Videntifier* che consente di interrogare la banca dati dei cosiddetti video illegali (sia nel caso di problemi di copyright che in presenza di immagini/video a carattere pedopornografico), senza la necessità di dover trasmettere in Rete nessuno dei dati sensibili coinvolti.

	Distanza dal Video A
Video A	0
Video *A* a cui è stata applicata una operazione di cambio di scala	0.0661
Video *A* a cui è stata applicata una operazione di rotazione	0.0787

Video *A* a cui è stata applicata una operazione di cropping	0.3794
Video *A* in cui la scena inquadrata è stata ripresa da un differente punto di vista	0.3819
Video B	0.5585

Tabella 3 – Distanze tra la rappresentazione del Video A e gli altri video considerati come esempio.



Video *A*

Video *A* a cui è stata applicata una operazione di cambio di scala

Figura 11 – Rappresentazione mediante hash percettivo.

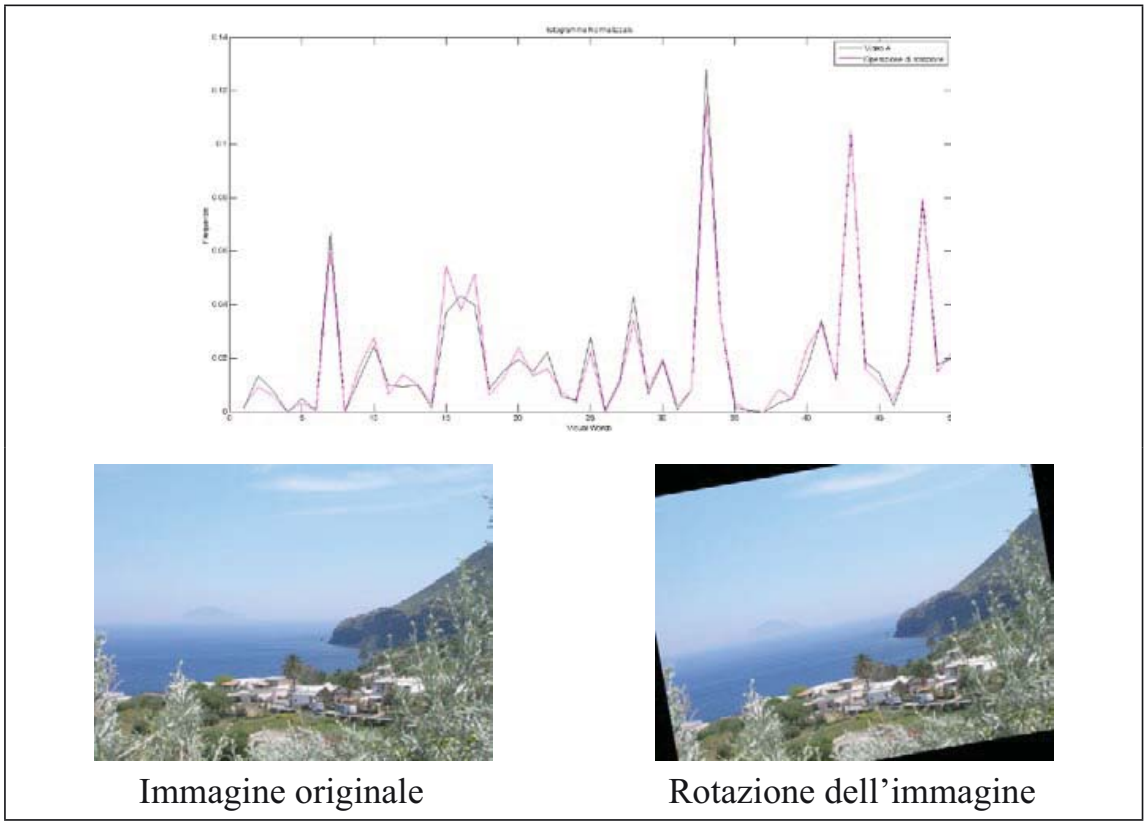


Figura 12 – Rappresentazione mediante hash percettivo.

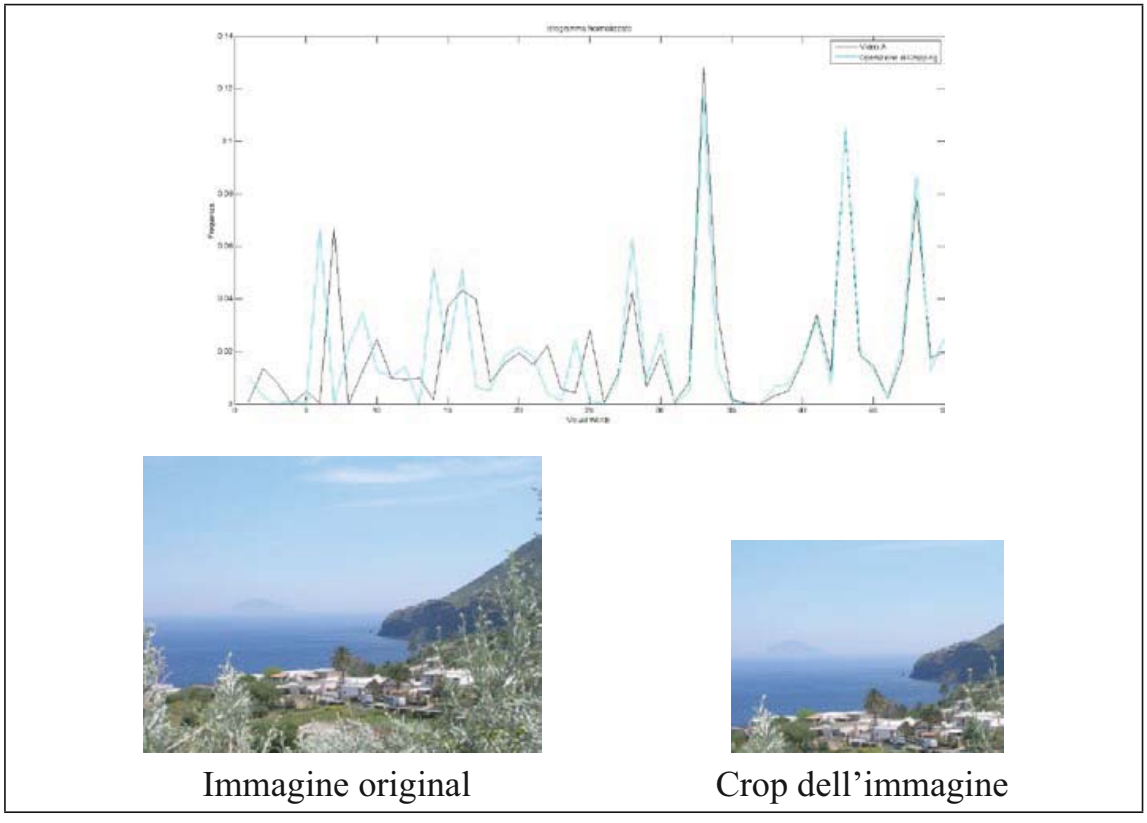


Figura 13 – Rappresentazione mediante hash percettivo.

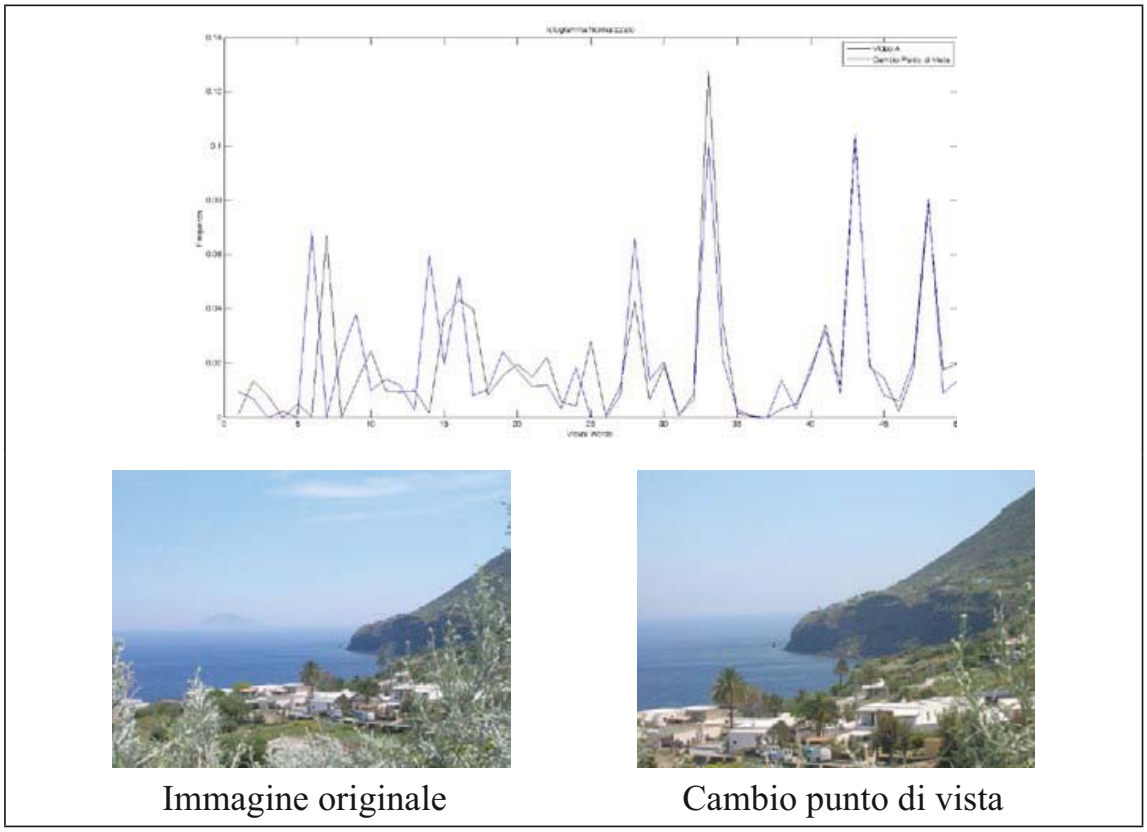


Figura 14 – Rappresentazione mediante hash percettivo.

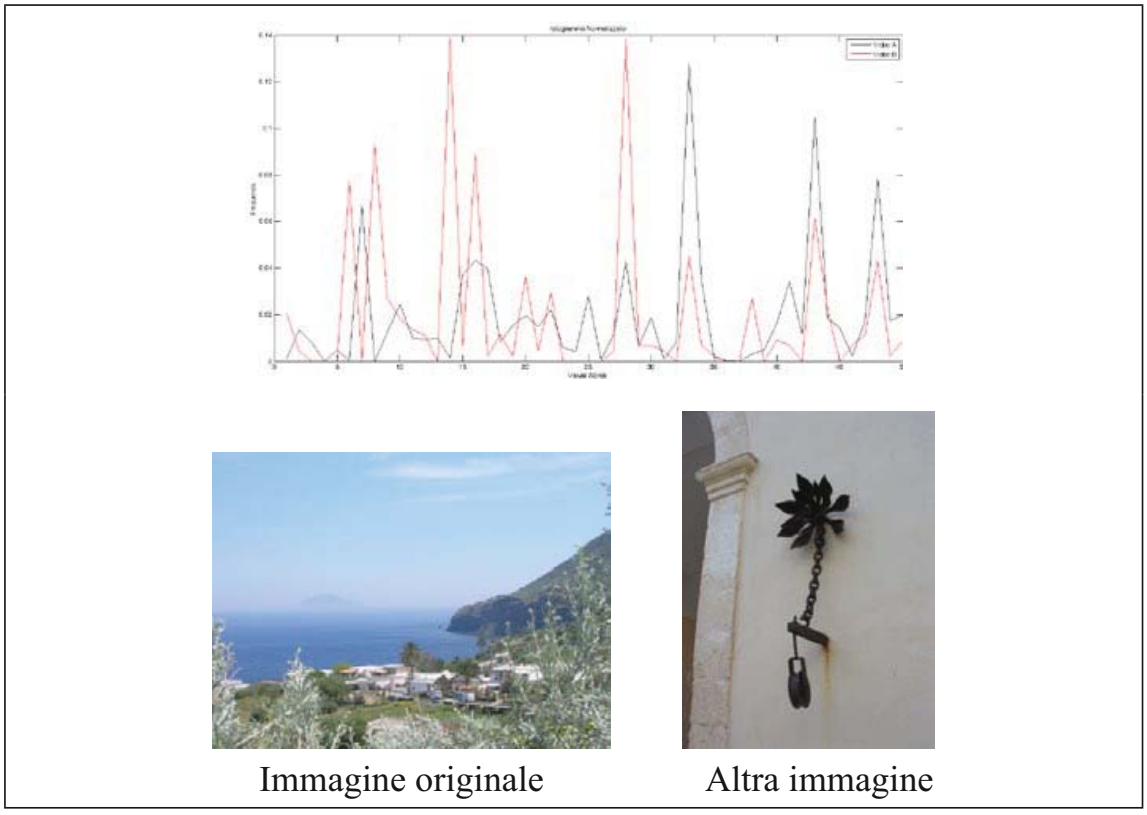


Figura 15 – Rappresentazione mediante hash percettivo.

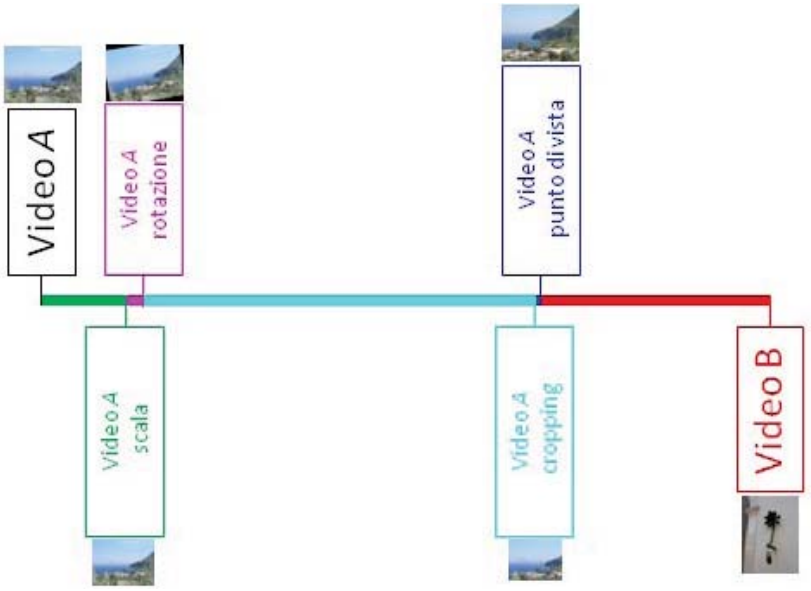


Figura 16 – Distanza dei vari video rispetto al video A in base all'hash percettivo.

5. SOFTWARE ESISTENTI

Concludiamo la nostra trattazione con un elenco di risorse e software dedicati all'analisi di immagini e video in ambito forense che pensiamo possano essere di interesse:

- AMPED Five (http://ampedsoftware.com)
- dTective di Avid e Ocean Systems ClearID (http://www.oceansystems.com/dtective/form/index.html http://www.oceansystems.com/dtective/clearid/proposal/index.html)
- Color Deconvolution (http://www.4n6site.com)
- LucisPro (http://www.LucisPro.com)
- Ikena Reveal di MotionDSP (http://www.motiondsp.com/products/IkenaReveal/form)
- Video Focus di Salient Stills (http://www.salientstills.com/products/videofocus/purchase.html)
- Impress di Imix (http://www.imix.nl/impress/impress_contact.htm)
- Video Investigator di Cognitech (http://www.cognitech.com/content/blogcategory/24/28/)
- StarWitness di SignalScape (http://www.starwitnessinfo.com/)
- VideoAnalyst di Intergraph (http://www.intergraph.com/learnmore/sgi/public-safety/forensic-video-analysis.xml)
- CrimeVision di Imagine Products (www.crimevision.net)

6. CONCLUSIONI

Le tecniche di *image* e *video forensics* costituiscono sicuramente un ulteriore strumento di indagine a disposizione degli investigatori per poter estrarre ed inferire, utili informazioni dalle immagini (e dai video) digitali. In questo documento ci si è soffermati su alcuni aspetti tecnici legati alle tecniche di rappresentazione e codifica dei video digitali. Sono stati anche introdotti alcuni concetti di base, necessari per poter comprendere i dettagli tecnici degli algoritmi presentati. Per essere in grado di recuperare o di inferire delle evidenze di prova è comunque necessaria una adeguata competenza specifica che richiede uno studio sistematico dei fondamenti della teoria dell'elaborazione delle immagini e dei video digitali. Gli stessi software oggi esistenti, che in qualche modo cercano di aiutare o quantomeno agevolare il lavoro degli investigatori, non riescono per forza di cose ad automatizzare in maniera sistematica ed efficiente tali operazioni e richiedono l'ausilio di utenti esperti.

7. BIBLIOGRAFIA

[1] J. F. Gantz, *The Diverse and Exploding Digital Universe - An Updated Forecast of Worldwide Information Growth Through 2011*, IDC White Paper, March 2008.
[2] Special Issue on Digital Forensics, Signal Processing Magazine, IEEE, Volume 26, Issue 2, March 2009.
[3] G. Reis, *Analisi Forense con Photoshop*, Apogeo - 2008 ISBN: 9788850327447.
[4] R. C. Gonzalez, R. E. Woods, *Elaborazione delle Immagini Digitali* - Terza edizione - ISBN: 9788871925066 Pearson - Prentice Hall Italia, Ottobre 2008.
[5] M. Chen, J. Fridrich, and M. Goljan, *Digital Imaging Sensor Identification (Further Study)*, Proceedings of the SPIE International Conference on Security, steganography, and Watermarking of Multimedia Contents IX, E. J. D. III and P. W. Wong, Eds., vol. 6505, no. 1. SPIE, 2007.
[6] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, *Determining Image Origin and Integrity Using Sensor Noise*, IEEE Transactions on Information Forensics and Security, volume 3, no.1, pp.74-90, March 2008.
[7] S. Battiato, G. Messina, *Digital Forgery Estimation into DCT Domain - A Critical Analysis*, Proceedings of ACM Multimedia 2009, Multimedia in Forensics (MiFor'09), Beijing, China, October 2009.
[8] W. Wang and H. Farid, *Exposing Digital Forgeries in Video by Detecting Double Quantization*, Proceedings of ACM Multimedia and Security Workshop, Sept. 2009, Princeton NJ.

[9] Wikipedia, *Scansione Interlacciata,* http:// it.wikipedia.org/wiki/Scansione_interlacciata.
[10] R. C. Gonzalez, R. E. Woods, op. cit., p. 2.
[11] S. Calderara, A. Prati, R. Cucchiara, *Trajectory Analysis in Video Surveillance for Multimedia Forensic*, Proceedings of ACM Multimedia 2009, Multimedia in Forensics workshop (MiFor), Beijing, China, October, 2009.
[12] G. Messina, S. Battiato, M. Mancuso, A. Buemi, *Improving Image Resolution by Adaptive Back-Projection Correction Techniques*, IEEE Transactions on Consumer Electronics, vol.48 n. 3 pp.409 -416, Agosto 2002.
[13] R. Rivest, The MD5 Message-Digest Algorithm, RFC Editor, 1992.
[14] H. Lejsek, F. H. Ásmundsson, K. Daðason, Á. Þór Jóhannsson, B. Þór Jónsson, L. Amsaleg, *Videntifier Forensic: A New Law Enforcement Service for Automatic Identification of Illegal Video Material*, Proceedings of ACM Multimedia 2009, Multimedia in Forensics workshop (MiFor), Beijing, China, October, 2009.
[15] S. Battiato, G. M. Farinella, G. C. Guarnera, T. Meccio, G. Puglisi, D. Ravì, R. Rizzo, *Bags of Phrases with Codebooks Alignment for Near Duplicate Image Detection*, Proceedings of ACM Multimedia 2010, Second ACM International Workshop on Multimedia in Forensics, Security and Intelligence (MiFor), Firenze, Italy, October, 2010.