UNIVERSITÀ DEGLI STUDI DI FIRENZE
Facoltà di Ingegneria - Dipartimento di Sistemi e Informatica

Tesi di Laurea Magistrale in Ingegneria Informatica

Annotazione automatica di video basata su co-occorrenza spazio temporale

Candidato
Gianluca Franchi

Relatori
Prof. Alberto Del Bimbo

Ing. Marco Bertini

Correlatori
Prof. Alberto Gandolfi

Ing. Lamberto Ballan

Ing. Giuseppe Serra

Anno Accademico 2008-2009

Alla mia famiglia

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

Il problema del riconoscimento automatico di contenuti video, sia eventi che oggetti (nel seguito indicati genericamente come concetti), ha ricevuto una grande attenzione nell'ambito della comunità scientifica.

La capacità da parte di un calcolatore di riconoscere un concetto all'interno di un'immagine o più genericamente di una sequenza video, grazie ad appositi algoritmi, risulta di notevole utilità nei più disparati settori applicativi [23, 6]. Gran parte di questo lavoro di analisi è svolto dai cosiddetti *concept detector* [7, 4, 3, 10].

Estraendo alcune caratteristiche peculiari, comunemente dette *feature*, come ad esempio colore, forma, o presenza di trame caratteristiche, e grazie anche al supporto di svariate tecniche di analisi, come la segmentazione o l'analisi spettrale, i *concept detector* riescono a dedurre il contenuto informativo di una sequenza. Ciascuna *feature* viene valutata tramite classificatori, tipicamente *Support Vector Machines* (SVMs) che forniscono una misura per la determinata caratteristica. In seguito tramite una fusione di basso livello che integra le informazioni tra le diverse *feature* si ottiene un valore che può essere inteso come la confidenza che il concetto sia presente nella scena. Tale quantità rappresenta una stima dell'accuratezza con cui i concetti sono riconosciuti ed è dipendente dall'attinenza delle caratteristiche della scena al modello conosciuto da identificare. Ad esempio, una batteria di *detector* potrebbe identificare all'interno di una immagine i seguenti concetti con le

relative confidenze indicate tra parentesi: automobile (35%), strada (22%), albero (16%) e vegetazione (8%) più una serie di altri concetti con probabilità molto basse.

Figura 1.1: Esempi di immagini video contenenti sport

Figura 1.2: Esempi di immagini video contenenti automobili

Purtroppo, essendo questo un campo in cui non sono state presentate soluzioni sufficientemente consolidate, i *detector* possono compiere molti errori, talvolta anche grossolani visto che, per loro natura, basano il proprio funzionamento su *feature* estratte dalle immagini senza avere alcuna conoscenza di informazioni di alto livello, non possedendo cioè una visione d'insieme della scena.

Ad esempio se consideriamo la *feature* istogramma di colore nello spazio di colore RGB, un'immagine con una forte presenza di verde può essere associata alla presenza di vegetazione, quando in realtà potrebbe semplicemente contenere un'autovettura di colore verde. È il cosiddetto problema del *gap semantico* introdotto da Smeulders *et al.* [24] , ovvero il divario che passa tra l'informazione che può essere estratta dai dati e l'interpretazione di quell'informazione da parte di un utente in un determinato contesto. Questo problema è difficilmente eliminabile e, come già accennato, nasce dal fatto che le *feature* sono informazioni di basso livello, basate cioè soltanto su caratteristiche locali dell'immagine e svincolate totalmente dal contesto in cui queste sono inserite.

Figura 1.3: Funzionamento dei detector

Per un essere umano la classificazione è semplice perché il nostro cervello è in grado di associare informazioni di basso livello a quelle di alto livello, quali ad esempio la presenza contemporanea di concetti correlati fra loro. Il cervello umano, infatti, è dotato di una base di conoscenza (sterminata confrontata a quella di un elaboratore) creata dall'esperienza che riesce ad utilizzare in maniera del tutto automatica tramite operazioni, anche molto complesse, difficilmente riproducibili da un calcolatore. Tornando all'esempio fatto in precedenza, il *detector* rileva la presenza dell'automobile ma non della strada. Un essere umano, avendo conoscenza del contesto, sa che un'automobile si sposta solitamente su una strada e che sono dunque molto rare le immagini in cui figuri la prima senza la seconda. Al contrario il *detector* analizza solo le *feature* dell'immagine e potrebbe facilmente attribuire al concetto di strada una percentuale molto bassa perché ad esempio l'immagine è stata scattata con un'angolazione tale da nasconderla in gran parte, rendendo poco visibili proprio quelle caratteristiche che ne avrebbero reso possibile il riconoscimento. In sostanza la strada è presente nell'immagine ma il *detector* non è in grado di riconoscerla perché la sola analisi delle *feature* non fornisce elementi sufficienti per effettuare il riconoscimento.

Se il *detector* potesse sfruttare informazioni legate al dominio, così com'è per gli esseri umani, potrebbe incrementare la percentuale relativa al concetto strada, anche senza riconoscerlo all'interno dell'immagine, semplicemente perché si tratta di un concetto fortemente correlato a quello di automobile riconosciuto con una percentuale alta (35%). Un meccanismo di questo tipo è molto interessante perché sfrutta oltre che le normali *feature*, su cui lavorano i *detector*, informazioni di più alto livello, quali ad esempio il le-

game statistico presente fra i concetti. Le relazioni statistiche tra i concetti possono essere sia di mutua occorrenza ma anche di mutua esclusione e presenza/assenza. Mediante tecniche di questo tipo sarebbe quindi possibile migliorare le probabilità di individuazione di concetti all'interno di *shot* video o immagini.

L'obiettivo di questo lavoro di tesi è appunto quello di effettuare una rielaborazione delle confidenze trovate dai *detector* specificatamente all'analisi di *shot* video, basandosi su informazioni di alto livello, quali ad esempio la correlazione tra concetti. Una ulteriore informazione che può essere utilizzata per migliorare le confidenze trovate dai *detector* è la continuità temporale. La coesione semantica di un video porta a mostrare una continuità temporale sia per quanto riguarda il contenuto visuale che per il contenuto semantico. Ritornando all'esempio precedente se in uno *shot* appare una macchina è altamente probabile che appaia anche negli *shot* dell'intorno temporale. È altrettanto probabile che nell'intorno temporale di quello *shot* sia presente una strada.

Il lavoro proposto in questa tesi prevede quindi due elaborazioni delle confidenze estratte dai *detector*:

- temporale
- relazionale

L'approccio temporale prevede di migliorare le probabilità di un concetto in uno *shot* analizzando il suo intorno temporale. Questo perché se un concetto è presente in uno *shot* è probabile che sia presente anche negli *shot* temporalmente vicini. Per ogni concetto viene quindi effettuato un raffinamento simile ad uno *smooth* sulla base del lavoro proposto da Liu *et al.* [18]. Per ogni *shot* si calcola un nuovo valore di confidenza della presenza del concetto in analisi come combinazione pesata delle probabilità di trovare il concetto negli *shot* dell'intorno. Queste probabilità vengono estratte da un insieme di *training* e sono quindi dipendenti dal concetto mentre il peso è determinato da una funzione pesudo-gaussiana per dare agli *shot* temporalmente più lontani un peso inferiore a quelli vicini.

L'approccio relazionale prevede invece di effettuare una fusione di probabilità all'interno del singolo *shot* sulla base delle relazioni di co-occorrenza, co-assenza o presenza incrociata. Per questo viene sfruttata la teoria dei modelli grafici in modo da apprendere tramite l'insieme di *training* un modello che contenga per ogni concetto parametri che si riferiscono alle possibili correlazioni tra i concetti. I risultati ottenuti mostrano che entrambi gli approcci portano ad un miglioramento in termini di performance sebbene il maggior contributo all'aumento delle prestazioni sia dovuto al raffinamento temporale.

1.1 Analisi preliminare

Il recupero e la classificazione di contenuti multimediali secondo la loro semantica è attualmente una delle sfide più difficili nella Computer Vision. L'estrazione di caratteristiche di alto livello mira a determinare la presenza o assenza di ogni concetto semantico come *Vegetazione*, *Esterno*, *Automobile*, ecc. nelle immagini o negli *shot* video. Il recente proliferare sul web di archivi digitali per contenuti multimediali, come *flickr* [1] o *YouTube* [2], ha reso necessario un approccio di ricerca più efficiente rispetto alla semplice annotazione manuale. Anche la digitalizzazione di archivi video preesistenti necessita una automatizzazione per il recupero dei video secondo il loro contenuto che attualmente può avvenire solo tramite la visione dell'intero archivio da parte di una figura umana.

L'approccio più comune per l'annotazione automatica di concetti è quella di utilizzare dei classificatori, tipicamente *Support Vector Machines* (SVMs), per estrarre la rilevanza tra immagini o video shot e un determinato concetto. I classificatori definiscono un modello a partire da caratteristiche di basso livello quali il colore, il contorno o la tessitura dell'immagine. Purtroppo esiste un problema non ancora risolto che riguarda il *gap* semantico tra le infor-

[1] www.flickr.com

[2] www.youtube.com

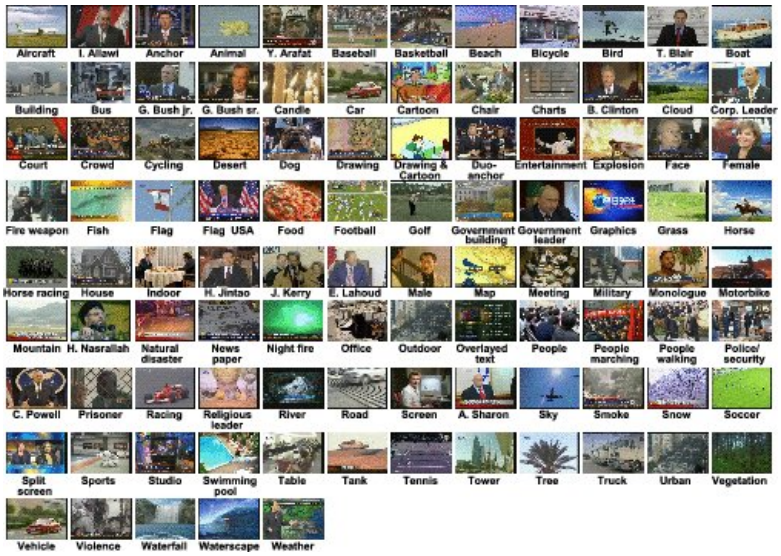

Figura 1.4: Esempi di immagini video contenenti il concetto automobile

mazioni di basso livello estratte dai classificatori e l'effettiva comprensione semantica della scena [24].

Un grande sforzo di ricerca è stato compiuto negli ultimi anni per costruire un gran numero di classificatori [25, 3, 10, 4] di diverse tipologie per esempio riferiti alla persona (faccia, mezzo busto, persona che corre, persona che cammina, ecc.), a oggetti (edificio, automobile, sedia, ecc.), a luoghi (montagna, esterni, interni, studio, ecc) e altre tipologie come sport, animali ecc.

Si può capire quanto sia grande il *gap* semantico da colmare pensando che i classificatori devono essere in grado di riconoscere ciascun concetto a partire da caratteristiche di basso livello. Per questo motivo si rendono necessarie tecniche che estraggano informazioni sui video di diversa natura, come la proprietà di persistenza temporale e la relazione di occorrenza tra concetti.

La relazione temporale si riferisce al fatto che tra gli *shot* di uno stesso video esiste spesso una coerenza per cui se un concetto è presente ad un istante temporale allora esiste una probabilità che lo stesso concetto sia presente anche agli istanti temporali vicini.

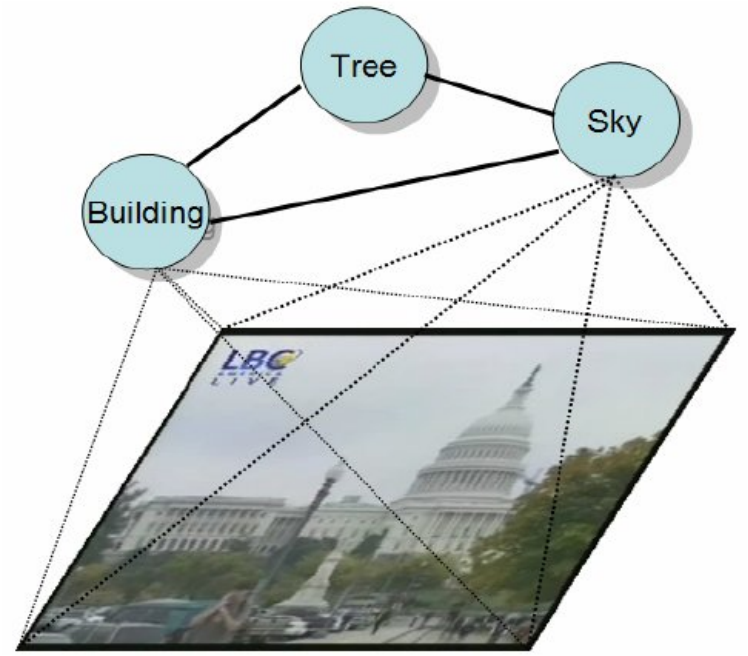

Figura 1.5: Rappresentazione tramite parole chiave del contenuto semantico di un frame

Le proprietà di occorrenza sono invece di "natura fisica" e si riferiscono al fatto che due o più concetti possono avere una alta probabilità di essere presenti contemporaneamente in uno *shot* oppure possono essere mutuamente esclusivi. Un esempio di co-occorrenza può essere: *montagna-esterno*; un esempio di mutua esclusione può essere: *interno-esterno*.

Possono esistere anche correlazioni incrociate come ad esempio la presenza di una *automobile* può occorrere con alta probabilità con il concetto *strada*, ma non è altrettanto probabile che se in uno *shot* si ha una *strada* vi sia anche una *automobile*.

1.2 Obiettivi

Questo lavoro di tesi mira a modificare i valori delle confidenze estratte dai *concept detector* in modo da aumentare i valori dei concetti effettivamente presenti nello *shot* e nello stesso tempo diminuire quelli dei concetti non presenti. Più precisamente, viene sviluppata un'applicazione in grado di elaborare le confidenze sfruttando le co-occorrenze spazio temporali al fine di migliorare la precisione di un sistema di annotazione automatica.

Il recupero dei contenuti multimediali e l'annotazione automatica avvengono tramite richieste a basi di dati che possono raggiungere dimensioni molto grandi. Le ricerche avvengono effettuando richieste di video o immagini che contengano una o più parole chiave immesse dagli utenti. Il recupero dei contenuti viene mostrato all'utente ordinato per rilevanza nello stesso modo in cui vengono presentati i risultati di una ricerca tramite il motore di ricerca *Google*[3].

La precisione nella rilevanza dei contenuti ritrovati è l'obiettivo di tutti i sistemi di recupero e di annotazione automatica. Per sistemi che restituiscono una sequenza ordinata di documenti è opportuno considerarne anche l'ordine in cui vengono presentati. In ambito scientifico si utilizza la cosiddetta *Average Precision* (AP) e *Mean Average Precision* (MAP) (che saranno definite con precisione nel capitolo 5) ovvero misure per valutare le prestazioni del sistema che enfatizzano la precisione nel presentare i risultati pertinenti in una posizione più rilevante.

1.3 Organizzazione della tesi

Nel prossimo capitolo verrà descritto lo stato dell'arte per quanto riguarda il raffinamento delle confidenze dei *concept detector*. Nel terzo capitolo verrà presentata la teoria sui modelli grafici utilizzati.

Nel quarto capitolo verranno analizzate le elaborazioni temporali e relazionali sviluppate nella tesi.

[3] www.google.com

Nel quinto capitolo verranno presentati i risultati ottenuti sugli insiemi di confidenze utilizzati come test per valutare le elaborazioni e infine verranno valutate le conclusioni finali e proposti alcuni sviluppi per migliorare il lavoro fin qui svolto.

Capitolo 2

Stato dell'arte

2.1 Concetti in immagini e video

Lo stato dell'arte dei classificatori o *concept detector* di concetti di alto livello a tutt'oggi non è soddisfacente. Nel lavoro di Yang *et al.* [30] si dimostra come buona parte dei classificatori utilizzi un approccio non affidabile poiché vengono costruiti in modo tale da non generalizzare, ovvero di non risultare precisi in video di dominio diverso rispetto a quelli utilizzati per il loro apprendimento. Anche nel lavoro di Aly e Hiemstra [1] viene evidenziato come solo i classificatori di concetti più frequenti e più chiaramente caratterizzati, consentono di apprendere modelli che risultino sufficientemente affidabili. Inoltre, le performance medie dei sistemi che partecipano a competizioni internazionali (come TRECVID) mostrano come lo sviluppo di *concept detector* efficienti non sia ancora giunto a un livello soddisfacente. Il progetto *MediaMill* [7] che è risultato uno dei sistemi più performanti di TRECVID 2009, mostra come i risultati di ricerche sui database offrono in media una bassa precisione. La *Mean Average Precision (MAP)* è una misura adottata nella competizione e viene calcolata facendo la media delle *Average Precision (AP)* definite precisamente nel capitolo 5. Le prestazioni in termini di MAP del progetto *MediaMill* rimangono al di sotto di 0.25 e la stragrande maggioranza dei sistemi rimane sotto 0.2. Ovvero in una logica di *ranking* solo un

quinto dei risultati proposti come pertinenti è effettivamente corretto. Per questo motivo si rende necessario, per avere una confidenza più precisa, utilizzare altre caratteristiche che sono proprietà dei video, ma che non vengono estratte dai singoli *concept detector*.

Estrazione delle Confidenze

Esistono diversi insiemi di *concept detector* [7, 10, 4, 3] che estraggono le feature di basso livello e tramite una serie di elaborazioni restituiscono valori di confidenza per ciascun concetto. L'approccio generale descritto da Snoek

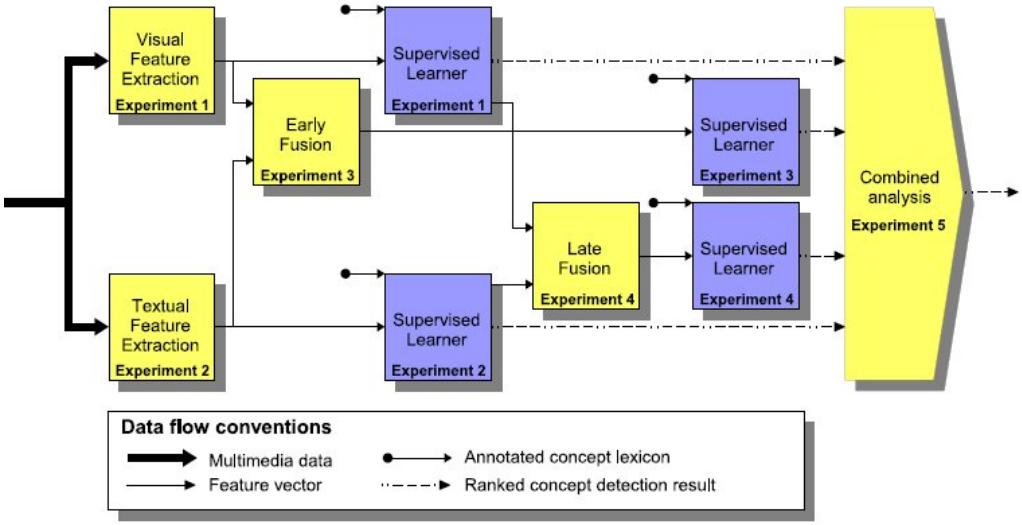

Figura 2.1: Generazione di confidenze da *feature* di basso livello

et al. [25] (mostrato in figura 2.1) prevede l'ulitizzo di classificatori *Support Vector Machines* (SVMs) per ciascun insieme di *feature* di basso livello.

Le *feature* estratte per caratterizzare i vari concetti sono diverse:

- Istogramma di colore
- Tessitura o caratteristiche visuali
- Punti salienti, corner, SIFT

- Audio
- ecc.

I risultati dei classificatori SVM vengono aggregati a diversi passi per aumentarne l'affidabilità e infine vengono trasformati tramite una funzione sigmoidea come descritto nel lavoro di Platt [20] per passare dal dominio di uscita dei classificatori SVM nel dominio più trattabile delle confidenze $[0, 1]$.

Per analizzare in modo approfondito il comportamento dei diversi *concept detector* sono stati creati degli istogrammi di frequenza delle confidenze. Si campiona cioè lo spazio di probabilità $[0, 1]$ con diecimila intervalli di larghezza costante; tale procedura è dovuta al fatto che utilizzando la precisione *double* si ha uno spazio troppo denso da trattare. Per ciascun concetto viene creato un istogramma di presenza delle confidenze che permette di visualizzare il comportamento del *concept detector*

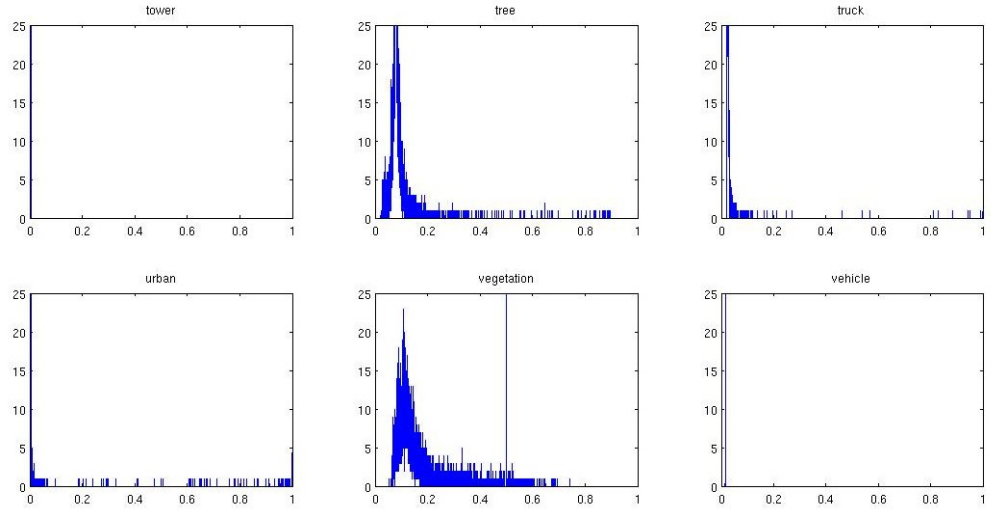

Figura 2.2: Istogrammi di presenza delle confidenze nell'intervallo di probabilità

In figura 2.2 vengono mostrati alcuni istogrammi da cui si nota come i comportamenti siano molto diversi in base al particolare *concept detector*. L'asse delle ordinate è stato limitato a 25 in quanto non è interessante vedere

quanti *shot* hanno la stessa confidenza ma è più importante vedere come si distribuiscono le confidenze nello spazio di probabilità.

Gli istogrammi di *tree* e *vegetation* mostrano un istogramma quasi ideale poichè si ha un picco di valore vicino allo zero giustificato dal fatto che nella maggior parte degli shot il concetto non è presente e si hanno diversi *shot* che presentano confidenze più alte che, nel caso ideale, identificano gli *shot* con effettiva presenza del concetto. Anche *urban*, e in modo più limitato *truck*, presentano un istogramma che può caratterizzare correttamente la probabilità di presenza del concetto. Sopratutto *urban* mostra, oltre al picco vicino al valore zero, anche una densità di presenze vicino a uno. Questo si avvicina al comportamento ideale di un sistema di annotazione automatica. Se si guardano invece gli istogrammi *tower* e *vehicle*, si nota come non siano presenti confidenze al di fuori di un ristretto *range* di valori vicino a zero. Questo significa che tutti gli *shot* vengono valutati con confidenze simili tra loro e risulta quindi molto difficile valutarne l'effettiva presenza.

Un'analisi basata su co-occorrenze spazio temporali consente di modificare questi valori di confidenza al fine di aumentare le confidenze in cui è effettivamente presente il concetto e abbassare quelle in cui il concetto non è presente.

2.2 Correlazione temporale

La correlazione temporale si riferisce al fatto che se in un video viene inquadrato un concetto è probabile che venga ripreso anche negli istanti precedenti e successivi. Siccome tra gli *shot* di uno stesso video esiste una continuità temporale si ha un'alta probabilità che uno stesso concetto venga inquadrato in istanti che stanno in una finestra temporale più o meno grande. In un video che riguarda una partita di calcio come mostrato in figura 2.3 questo fenomeno è altamente presente in quanto per diverse sequenze di *shot* saranno presenti gli stessi concetti

Nel lavoro di Yang e Hauptmann [31] viene analizzata la consistenza temporale definita rispetto ai concetti semantici discutendo le implicazioni all'a-

Figura 2.3: Video con alta correlazione temporale

nalisi video per il recupero dei contenuti multimediali. Questo viene mostrato calcolando i risultati ottenuti tramite fusione lineare e fusione ad oracolo (utilizzando concetti etichettati manualmente negli *shot* adiacenti) delle probabilità della presenza di un concetto data la sua presenza negli *shot* precedenti. Viene dimostrato, tramite l'approccio interattivo, come l'utilizzo della consistenza temporale apporti un effettivo miglioramento alle prestazioni del sistema. Tuttavia, a causa dell'alta variabilità degli *shot* nei video, in generale non risulterebbe conveniente ricalcolare le confidenze tramite combinazione lineare dei valori di confidenza degli *shot* precedenti.

Nell'articolo di Liu *et al.* [18] viene mostrato un approccio simile, in cui viene proposto un filtro che sfrutta la dipendenza temporale tra *shot*. Tramite un insieme di *train* vengono calcolate per ciascun concetto le probabilità condizionate della presenza/assenza del concetto data l'effettiva presenza o assenza del concetto negli *shot* precedenti e successivi. La probabilità della presenza di un concetto viene quindi calcolata facendo la combinazione lineare delle probabilità condizionate calcolate sull'insieme di *train* con le confidenze ottenute dai detector pesate da un valore che assegna il peso in base alla distanza temporale. Il peso che viene stimato tramite test *chi_quadro* (χ^2) ha la funzione di dare maggior importanza agli *shot* temporalmente vicini e minore a quelli più distanti. L'approccio usato in questa tesi ricalca questo articolo e verrà descritto in modo più approfondito nel capitolo 4.

2.3 Relazione di occorrenza

La relazione di occorrenza si riferisce all'informazione che è possibile inferire dalle relazioni che intercorrono tra due diversi concetti. Queste relazioni possono essere del tipo di co-occorrenza, mutua esclusione (se è presente uno l'altro è assente) o relazione di co-assenza (se non è presente uno non è presente neppure l'altro).

Le relazioni di co-occorrenza si riferiscono alla probabilità che due concetti siano presenti contemporaneamente nello stesso *shot* video. Possono essere dovute al fatto che un concetto specifica un altro quindi se è presente il concetto *cane* è presente anche il concetto *animale*. Oppure possono essere dovute a relazioni di appartenenza, per esempio se è presente un *anchorman* è presente sicuramente anche il concetto *faccia*.

Le relazioni di mutua esclusione esprimono la probabilità che data la presenza di un concetto si abbia l'assenza dell'altro. Un esempio di mutua esclusione sono le coppie *desert-indoor*, *tennis-camion* ecc.

Le relazioni di co-assenza sono simili alle relazioni di co-occorrenza, ma vengono rilevate in modo statistico quando entrambi i concetti sono assenti contemporaneamente. Un esempio di coppie di concetti che possono avere una relazione di assenza diversa da zero è: *cielo-nuvola* oppure *urban-building*. Quando si intende utilizzare la correlazione di occorrenza è necessario fondere le confidenze dei concetti con l'informazione che deriva dalle relazioni con gli altri concetti.

In letteratura sono stati presentati numerosi tentativi di sfruttare questo tipo di informazione, in gran parte sfruttando relazioni di mutua occorrenza modellate attraverso tecniche classiche della teoria dell'informazione. Sono state studiate anche altre tecniche di annotazione automatica basate su diversi aspetti importanti. Per esempio, Liu *et al.* [17] sfrutta le relazioni tra parole e immagini, Zha *et al.* utilizza le ontologie [34], mentre Zeng *et al.* [36] e Tian *et al.* [26] propongono un *ranking* ottimizzato dei valori di confidenza. Nel lavoro di Zheng *et al.* [35] viene utilizzata una tecnica chiamata *Pointwise mutual information weighted scheme* che sfrutta il concetto di mutua informazione. La mutua informazione puntuale tra due concetti può essere vista

come la quantità di informazione che un concetto contiene rispetto all'altro; se due concetti sono indipendenti hanno mutua informazione pari a zero. La mutua informazione puntuale normalizzata viene utilizzata come peso per stimare la probabilità di un concetto in uno *shot* sulla base delle informazioni di tutti gli altri concetti a disposizione. Sia d uno *shot* e $p(X_i(d) = x_i)$ la confidenza associata al concetto i nello *shot* d, si ha che la probabilità associata ad un concetto y viene calcolata come combinazione linare rispetto a tutti gli altri concetti

$$p(Y(d) = y) = \sum_i weight_{C_i} * p(X_i(d) = x_i)$$

il peso viene determinato dalla mutua informazione

$$p(Y(d) = y) = \sum_i \alpha \log \frac{p(x_i|y)}{p(x_i)} * p(X_i(d) = x_i)$$

dove $p(x_i) = Mean_d(p(X_i(d) = x_i))$ e $p(x_i|y)$ viene approssimato come la confidenza del concetto i in uno *shot* di *query* predefinito.

Nel lavoro di Kennedy *et al.* [11] viene proposto un approccio di fusione delle confidenze di concetti che hanno una relazione sia positiva che negativa tramite un riordinamento automatico a due fasi. Le relazioni vengono estratte da una prima ricerca e vengono analizzati gli *shot* che hanno le confidenze maggiori e minori. Analizzando questi *shot* vengono estratti i concetti con le confidenze maggiori indicandoli come concetti pseudo positivi e pseudo negativi. In una seconda fase di ricerca del concetto la lista dei risultati viene riordinata aumentando la preferenza degli *shot* che presentano una alta confidenza dei concetti pseudo positivi e diminuendo quelli che presentano una alta confidenza dei concetti pseudo negativi. Questo approccio ottiene piccoli risultati per quanto riguarda l'aumento di prestazioni, ma prevede una fusione automatica senza alcuna conoscenza di informazioni a priori.

Nell'articolo di Wei *et al.* [28] viene studiata una strategia di fusione di concept detector che si basa su quattro aspetti fondamentali per il video retrieval:

- *Semantica*: si riferisce alla correlazione tra concetti, sfrutta le strutture di ontologia per costruire uno spazio vettoriale semantico nel quale individuare concetti semanticamente vicini

- *Osservabilità*: si riferisce alla estrazione di concetti anche semanticamente non correlati, ma che possono avere una relazione in un dominio video. Viene costruito uno spazio di osservabilità per estrarre concetti tramite correlazione di concetti con vista globale.

- *Affidabilità*: si riferisce alla robustezza dei detector, che viene aumentata tramite la fusione congiunta di un insieme di detector correlati utilizzando la tecnica descritta nel lavoro di Kennedy *et al.* [11].

- *Diversità*: si riferisce alla varietà di detector utilizzati.

Vengono creati quattro sotto insiemi di concetti $\mathcal{A}, \mathcal{B}, \mathcal{P}, \mathcal{N}$. Dato un concetto X, si estraggono

- $\mathcal{A}$ è l'insieme dei concetti più vicini estratti dallo spazio semantico al concetto X

- $\mathcal{B}$ è l'insieme dei concetti più vicini estratti dallo spazio di osservabilità a ciascun concetto nell'insieme $\mathcal{A}$

- $\mathcal{P}$ e $\mathcal{N}$ sono gli insiemi dei concetti rispettivamente più pertinenti e meno pertinenti estratti con il metodo descritto nel lavoro di Kennedy *et al.* [11] per ogni concetto in $\mathcal{A}, \mathcal{B}$.

La fusione viene effettuata tramite una serie di funzioni, ognuna delle quali apporta il contributo di un insieme di concetti. Per migliorare l'affidabilità del sistema per ogni concetto viene ricalcolato il valore di confidenza fondendo i concetti più e meno pertinenti tramite la seguente funzione:

$$\begin{aligned}\hat{D}(C) = D(C) &+ \frac{1}{|\mathcal{P}|} \sum_{P_i \in \mathcal{P}} Observability(P_i, C) * D(P_i) \\ &- \frac{1}{|\mathcal{N}|} \sum_{N_i \in \mathcal{N}} Observability(N_i, C) * D(N_i)\end{aligned} \tag{2.1}$$

dove C è un concetto in $\mathcal{A}, \mathcal{B}$, $D(C)$ è il valore di uscita del *concept detector* C e $Observability(N_i, C)$ è una misura di distanza tra i concetti N_i e C calcolato nello spazio di osservabilità.

Per migliorare l'osservabilità dei concetti correlati semanticamente viene effettuata una fusione simile a 2.1. Per ogni concetto nell'insieme $\mathcal{B}$ viene estratto un sotto insieme di $\mathcal{A}$ chiamato $N(\mathcal{A})$ che contiene i concetti di A più vicini nello spazio di osservabilià. Da cui per tutti i concetti A dell'insieme $\mathcal{A}$ viene ricalcolato

$$\hat{D}(A) = D(A) + \frac{1}{|\mathcal{N}(\mathcal{A})|} \sum_{B_i \in \mathcal{N}(\mathcal{A})} Observability(B_i, A) * D(B_i) \qquad (2.2)$$

La funzione di fusione per ottenere pertinenza del concetto di *query* X nello *shot* I si ottiene fondendo X con ciascun concetto dell'insieme $\mathcal{A}$:

$$Sim(X, I) = \sum_{A_i \in \mathcal{A}} Semantic(X, A_i) * Score(\hat{D}(A_i), I)$$

dove $Semantic(X, A_i)$ è un valore di distanza tra i concetti A_i e X calcolato nello spazio semantico e $Score(\hat{D}(A_i), I)$ è l'uscita con fusione del concetto A_i nello *shot* I.

I risultati di questo articolo mostrano come tutte queste caratteristiche concorrono a migliorare le prestazioni complessive del sistema di recupero, sebbene la fusione delle confidenze come combinazione lineare non sia "rigorosa".

Nel lavoro di Weng *et al.* [29] viene proposto un approccio che esplora ed integra sia le correlazioni di occorrenza sia quelle temporali tra *shot*. Utilizza un algoritmo ricorsivo in cui in un primo passo vengono apprese entrambe le relazioni da un insieme di valori di verità preannotati correttamente. Questo sistema non prevede un approccio globale per le correlazioni di occorrenza, ma estrae sottoinsiemi di concetti tramite un algoritmo di partizionamento come mostrato in figura 2.4.

Successivamente fonde le informazioni utilizzando una tecnica basata sui modelli grafici. La tecnica prevede la ricerca di un'assegnazione delle etichette (presenza-assenza di ciascun concetto) tale che sia minimizzata la distanza tra la probabilità dell'assegnazione calcolata sul modello con il valore del corrispondente *concept detector*. Inoltre prevede che la probabilità dell'assegnazione abbia distanza minima con le confidenze dei correlati e distanza minima

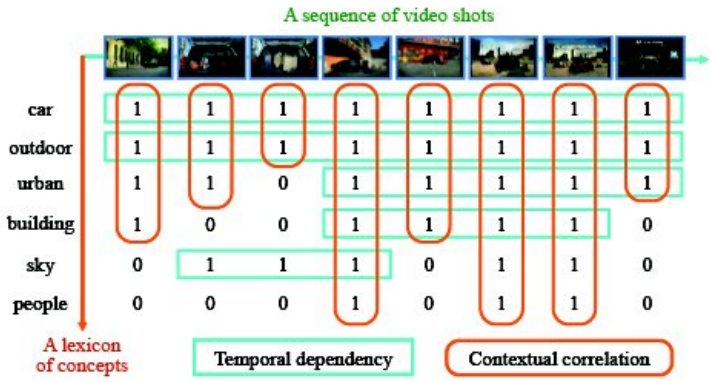

Figura 2.4: Tecnica di fusione nell'articolo di Weng *et al.* [29]

con le confidenze calcolate su *shot* temporalmente vicini. L'ottimizzazione viene effettuata tramite la tecnica dei minimi quadrati.

Oltre al lavoro di Weng *et al.* [29] esistono altri approcci che utilizzano i modelli grafici per effettuare fusioni di informazioni di co-occorrenza come dimostra il lavoro di Li [16]. L'articolo di Zha *et al.* [33] si concentra nello sfruttare le caratteristiche dei modelli grafici per ottenere un approccio di apprendimento semi supervisionato. L'approccio prevede di calcolare i valori di etichettatura presenza-assenza minimizzando una funzione di errore che considera la correlazione tra concetti. La soluzione viene calcolata risolvendo la funzione di errore in una equazione di *Sylvester* [12].

Il lavoro di Qi *et al.* [21] propone un metodo per etichettare *shot* con presenza-assenza di concetti. Utilizza un modello grafico di Gibbs estraendo direttamente dalle caratteristiche di basso livello le informazioni di relazione come mostrato in figura 2.5. Tramite l'approccio *Maximum A Posteriori* descritto al terzo capitolo, determina il vettore delle etichette che, tra tutti, massimizza la probabilità del vettore di etichette dato il modello. Per valutare la precisione di recupero dei contenuti ordina gli *shot* in base ad una funzione di aspettativa ovvero un valore che sostituisce l'osservazione del detector per un dato concetto. Questo tipo di approccio risulta essere promettente, ma molto oneroso dal punto di vista computazionale.

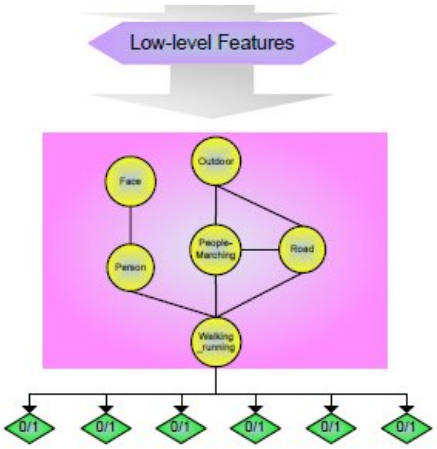

Figura 2.5: Tecnica di fusione dalle feature di basso livello

Un'altra tecnica che prevede di calcolare i valori di presenza assenza per ogni concetto viene proposta in da Wang *et al.* [27] in cui si descrive una tecnica di apprendimento di tipo trasduttivo (ovvero senza il calcolo della funzione di assegnazione delle etichette) tramite un *hidden Markov Random Field*. L'idea prevede di ottenere una possibilità di annotare più concetti nello stesso *shot*. Viene realizzato cercando il vettore di etichette che massimizza la funzione di energia del modello formata dal prodotto della probabilità di similarità con le confidenze dei *concept detector*, compatibilità tra *shot* con caratteristiche simili e consistenza tra concetti correlati sia in modo positivo, sia in modo negativo sia con correlazione incrociata.

Nell'articolo di Chen e Hauptmann [32] e in quello di Hauptmann *et al.* [9] viene proposto un modello grafico chiamato *Discriminative Random Field*. Il *Discriminative Random Field* è un modello che utilizza due termini fondamentali:

- l'associazione che descrive la relazione tra le osservazioni dei detector e le label per ogni singolo nodo;
- l'interazione che descrive il modello di co-occorrenza e correlazione tra due nodi.

La probabilità condizionale di avere un set di label Y data l'osservazione dei detector X si può scrivere utilizzando i modelli grafici come:

$$P(Y|X) = \frac{1}{Z} exp\Big(\sum_{i \in S} A_i(y_i, W, X) + \sum_{i \in S} \sum_{j \in N_i} I_{i,j}(y_i, y_j, V, X) \Big)$$

in cui X denota le osservazioni, Y denota le label, Z è una funzione di normalizzazione, A una funzione di associazione tra le osservazioni e le label, $I_{i,j}$ una funzione di correlazione tra le osservazioni dei concetti i e j. S è l'insieme dei concetti, V e W sono i parametri del modello rispettivamente di correlazione e di associazione. Uno dei problemi maggiori nel gestire i modelli grafici è il calcolo della normalizzazione Z che nel caso di grafo completamente connesso dev'essere effettuata su tutte le coppie di nodi possibili che sono 2^N per N nodi. Il calcolo della Z viene quindi approssimata.

Per utilizzare un modello più semplice ed evitare il peso del calcolo della normalizzazione viene creato anche un modello approssimato che consente di fattorizzare la probabilità

$$\begin{aligned} P(Y|X, W) &\simeq exp\Big(\sum_{i \in S} y_i W_{ii}^T u_{ii}(X) + \sum_{i \in S} \sum_{j \in N_i} y_i y_j W_{ij}^T u_{ij}(X) \Big) \\ &= \prod_{i \in S} exp\Big(y_i W_{ii}^T u_{ii}(X) \Big) \prod_{i \in S} \prod_{j \in N_i} y_i y_j W_{ij}^T u_{ij}(X) \end{aligned} \tag{2.3}$$

Questa approssimazione chiamata *Generalized DMRF* è più trattabile anche in fase di apprendimento dei parametri soprattutto quando l'insieme dei parametri è grande. I test sono effettuati con insiemi costituiti da dieci concetti. Negli sviluppi futuri sono indicate due evoluzioni importanti che vengono trattate in questo lavoro di tesi come il considerare le correlazioni incrociate oltre alle correlazioni positive/negative, e sfruttare le informazioni derivate dalla correlazione temporale.

In questo lavoro di tesi si intende effettuare una rielaborazione dei valori delle confidenze attraverso due raffinamenti: temporale e relazionale. Il raffinamento temporale si basa sul lavoro di Liu *et al.* [18]. Viene utilizzata una diversa funzione dei pesi assegnati alle probabilità condizionate e una diversa

finestra temporale. Il raffinamento relazionale viene effettuato proponendo un modello grafico che prende spunto dal DMRF di Chen e Hauptmann [32] e dal lavoro di Qi *et al.* [21]. Tale modello viene studiato appositamente per essere in grado di valutare tutti i tipi di relazione statistica possibili tra ciascuna coppia di concetti.

Capitolo 3

Modelli Grafici

3.1 Modelli grafici

I modelli grafici probabilistici [2, 15, 14] sono strumenti eleganti che combinano probabilità e struttura logica per rappresentare in modo compatto ed efficace complessi fenomeni del mondo reale. Esisono due tipologie di modelli grafici:

- Reti Bayesiane, rappresentate da un grafo direzionale aciclico;
- Reti di Markov, rappresentate da un grafo non direzionale.

Siano $\mathcal{X} = \{X_1, ..., X_n\}$ un insieme di variabili casuali, i modelli grafici offrono uno strumento per rappresentare la distribuzione di probabilità congiunta $P(X_1, ..., X_n)$. In questo lavoro di tesi sono stati studiati in particolar modo i modelli grafici non direzionali detti anche *Markov Random Field* o reti di Markov descritti in Kindermann [13]. Sono strumenti adatti a modellare una varietà di fenomeni dove non è possibile strutturare una direzionalità delle relazioni tra i nodi oppure dove le relazioni presentano strutture cicliche.

Nei modelli grafici ogni vertice rappresenta una variabile casuale e gli archi rappresentano le interazioni tra di essi. Per parametrizzare le reti di Markov si utilizza la definizione di *fattori* o *funzioni potenziali*.

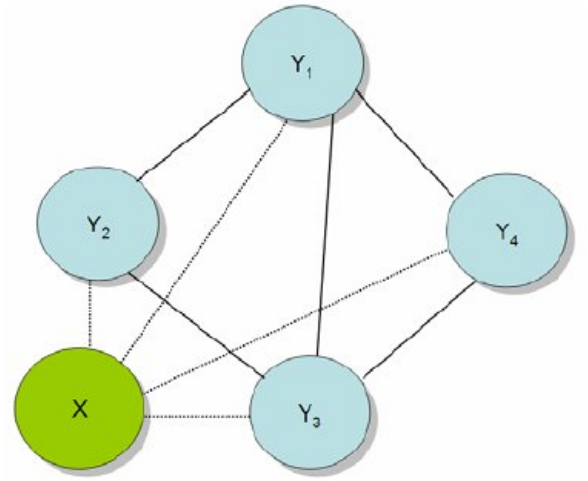

Figura 3.1: Esempio di grafo non direzionato che rappresenta una rete di Markov

Definizione 3.1.1 *Sia $\mathcal{G}$ una struttura di una rete di Markov. Una distribuzione P si dice che fattorizza $\mathcal{G}$ se si ha:*

- *Una serie di sottoinsiemi dei dati $D_1, ..., D_M$ dove ogni D_i è un sottografo completo di $\mathcal{G}$*
- *una serie di funzioni $\phi(D_i) : Val(D) \to \mathbb{R}^+$*

tali che:

$$P(X_1, ..., X_M) = \frac{1}{Z} P'(X_1, ..., X_M)$$

dove

$$P'(X_1, ..., X_M) = \phi(D_1) * \phi(D_2) * ... * \phi(D_M)$$

è una misura di probabilità non normalizzata e

$$Z = \sum_{X_1,...,X_M} P'(X_1, ..., X_M)$$

è una costante di normalizzazione chiamata funzione di partizione.

Una distribuzione P che fattorizza $\mathcal{G}$ è anche chiamata distribuzione di *Gibbs* da *Josiah Willard Gibbs* ingegnere, chimico e fisico statunitense.

L'unico vincolo sui parametri è la non negatività dei fattori (Teorema di Hammersley Clifford [5]) per cui è possibile utilizzare fattorizzazioni esponenziali che assicurano la non negatività.

Si consideri un set $\mathbf{X}$ di variabili aleatorie, un grafo $\mathcal{G}$ i cui nodi sono variabili in $\mathbf{X}$ e l'insieme $\mathcal{C}$ di tutte le cricche (massimali) in $\mathcal{G}$. In teoria dei grafi, una cricca (o *clique*) in un grafo non orientato $\mathcal{G}$ è un insieme V di vertici tale che, per ogni coppia di vertici in V, esiste un arco che li collega. In modo equivalente, si potrebbe dire che il sottografo indotto da V è un grafo completo.

Date le funzioni potenziali per le cricche in $\mathcal{C}$ la distribuzione di probabilità congiunta di $\mathbf{X}$ viene calcolata come segue:

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{x_c})$$

dove Z è la funzione di partizione e $\mathbf{x_c}$ è lo stato della cricca c quando $\mathbf{X} = \mathbf{x}$. Lo stato di una cricca c in un *Markov Random Field* è una specifica realizzazione delle variabili in c ovvero un evento definito da una specifica configurazione dei valori delle variabili in $\mathbf{X}$. La funzione di partizione è data dalla somma su tutte le possibili configurazioni:

$$Z = \sum_{\mathbf{x} \in \mathcal{X}} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{x_c})$$

È utile riscrivere le funzioni di potenziale sotto forma di esponenziali come:

$$\phi(\mathbf{x}) = exp(-E[\mathbf{x}])$$

dove $E[\mathbf{x}] = -\log \phi(\mathbf{x})$ è spesso chiamata *funzione di energia*. L'utilizzo della parola energia deriva dalla statistica fisica dove la probabilità di uno stato fisico dipende inversamente dalla sua energia.

La sottoclasse delle reti di Markov trattata in questa tesi è anche chiamata *pairwise Markov network* e rappresenta una distribuzione in cui i fattori

vengono presi come singole variabili o al più coppie di variabili. Più precisamente una *pairwise Markov network* su un grafo $\mathcal{G}$ è formato da un insieme di fattori chiamati *campi* sui singoli nodi $\phi(x_i) : \{i = 1, ..., M\}$ e da un insieme di fattori chiamati *interazioni* tra i nodi $\phi(x_i, x_j) : (x_i, x_j) \in \mathcal{G}$.

3.2 Indipendenza nelle Reti di Markov

Una nozione fondamentale dei modelli grafici è l'indipendenza condizionale.

Definizione 3.2.1 *Siano* $\mathbf{X}, \mathbf{Y}$ *e* $\mathbf{Z}$ *insiemi di variabili aleatorie. Si dice che* $\mathbf{X}$ *è condizionalmente indipendente a* $\mathbf{Y}$ *dato* $\mathbf{Z}$ *in una distribuzione* P *se*

$$P(\mathbf{X} = x, \mathbf{Y} = y|\mathbf{Z} = z) = P(\mathbf{X} = x|\mathbf{Z} = z)P(\mathbf{Y} = y|\mathbf{Z} = z)$$

per ogni $x \in Val(\mathbf{X}), y \in Val(\mathbf{Y}), z \in Val(\mathbf{Z})$

Verrà usata la comune notazione $\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}$ per dire che $\mathbf{X}$ è condizionalmente indipendente da $\mathbf{Y}$ dato $\mathbf{Z}$.

Si possono definire due assunzioni di indipendenza: la proprietà di Markov locale e la proprietà di Markov globale. La proprietà di Markov locale è associata ad ogni nodo nel grafo ed è basata sul fatto che è possibile definire tutte le influenze ad un nodo condizionandolo ai nodi adiacenti. La proprietà di Markov globale è associata a sottografi in cui si ha che un insieme di nodi può essere indipendente da un altro insieme di nodi dato un sottografo intermedio.

Definizione 3.2.2 *Sia* $\mathcal{G}$ *un grafo non diretto. Per ogni nodo* X *appartenente all'insieme dei nodi* $\mathcal{X}$*, sia* $\mathcal{N}_\mathcal{G}(X)$ *l'insieme dei nodi adiacenti a* X *nel grafo detto Markov blanket. Si definisce indipendenza locale di Markov associata ad* $\mathcal{G}$

$$\mathcal{I}_l(\mathcal{G}) = (X \perp \mathcal{X} - X - \mathcal{N}_\mathcal{G}(X)|\mathcal{N}_\mathcal{G}(X)) : X \in \mathcal{X}$$

Ciò significa che X è indipendente dal resto dei nodi nel grafo data la conoscenza delle istanze dei nodi adiacenti.

Per quanto riguarda la proprietà globale è necessario definire il *percorso attivo* su una rete di Markov e il concetto di *separazione*.

Definizione 3.2.3 *Sia $\mathcal{G}$ una rete di Markov e $X_1 - ... - X_k$ un percorso in $\mathcal{G}$, sia $\mathbf{E} \subseteq \mathcal{X}$ un sottoinsieme di variabili osservate. Si dice che il percorso $X_1 - ... - X_k$ è attivo dato $\mathbf{E}$ se nessuno dei nodi X_i, $i = \{1, 2, ..., k\}$ del percorso appartiene a $\mathbf{E}$.*

Definizione 3.2.4 *Si dice che un insieme di nodi $\mathbf{Z}$ separa $\mathbf{X}$ e $\mathbf{Y}$ in $\mathcal{G}$ e si scrive $sep_{\mathcal{G}}(\mathbf{X}; \mathbf{Y}|\mathbf{Z})$, se non esistono percorsi attivi tra ciascun nodo $X \in \mathbf{X}$ e $Y \in \mathbf{Y}$ dato $\mathbf{Z}$.*

A questo punto è possibile definire la proprietà globale di Markov.

Definizione 3.2.5 *Si definisce proprietà globale di Markov associata a $\mathcal{G}$*

$$\mathcal{I}(\mathcal{G}) = (\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}) : sep_{\mathcal{G}}(\mathbf{X}; \mathbf{Y}|\mathbf{Z})$$

Questo significa che se $\mathbf{Z}$ separa $\mathbf{X}$ e $\mathbf{Y}$ allora vale l'indipendenza condizionale $(\mathbf{X} \perp \mathbf{Y}|\mathbf{Z})$.

È quindi possibile effettuare una connessione tra la proprietà locale e globale.

Teorema 3.2.1 *Sia P una distribuzione su $\mathcal{X}$ e $\mathcal{G}$ una rete di Markov che rappresenta la struttura di connessione di $\mathcal{X}$, se P è una distribuzione di Gibbs su $\mathcal{G}$ allora valgono tutte le proprietà locali di Markov associate a $\mathcal{G}$.*

Si nota come il teorema non vale sempre in direzione contraria ma solo se la distribuzione è positiva. Questo è conosciuto come teorema di Hammerseley-Clifford dimostrato in Clifford [5].

Teorema 3.2.2 *Sia P una distribuzione positiva su $\mathcal{X}$ e $\mathcal{G}$ una rete di Markov che rappresenta la struttura di connessione di $\mathcal{X}$, se tutti i vincoli di indipendenza condizionata in $\mathcal{G}$ valgono in P allora P è una distribuzione di Gibbs su $\mathcal{G}$.*

Questo risultato mostra come per distribuzioni positive su $\mathcal{X}$ la proprietà globale di Markov implica che la distribuzione sia fattorizzabile in accordo con la struttura della rete di Markov. Per questa classe di distribuzioni si ha che una distribuzione P è fattorizzabile su una rete di Markov $\mathcal{G}$ se e solo se tutte le indipendenze condizionali evidenziate dalla struttura di $\mathcal{G}$ valgono in P. Nella figura 3.2 si nota come ciascun percorso da qualsiasi nodo A a qualsiasi nodo B passa attraverso un nodo C. Di conseguenza la proprietà di indipendenza condizionata $\mathbf{A} \perp \mathbf{B}|\mathbf{C}$ vale per ogni distribuzione descritta dal grafo.

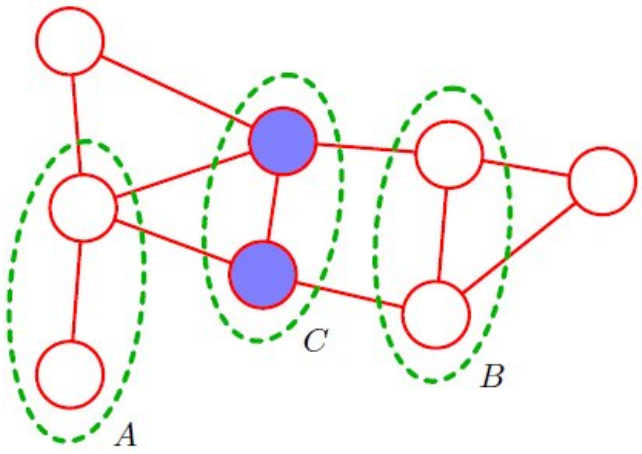

Figura 3.2: Esempio di grafo per cui vale l'indipendenza condizionata

Si denoti con $mb(X_i)$ lo stato del sottoinsieme di nodi collegati con X_i chiamato *Markov Blanket*. Data le probabilità di indipendenza condizionale nelle reti di Markov, è possibile calcolare la distribuzione di probabilità condizionata $P(X = x|mb(X))$ ovvero, la probabilità che un nodo X si trovi nello stato x data la conoscenza del rispettivo *Markov Blanket*. Sia $MB(X_i)$

l'insieme dei nodi collegati con X_i si ha che:

$$P(X = x|mb(X)) = \frac{P(X = x, MB(X) = mb(X))}{P(MB(X) = mb(X))}.$$

Siccome $P(MB(X) = mb(X))$ rimane costante per ogni valore di X, si può ignorare questo termine ottenendo la seguente equazione:

$$P(X = x|mb(X)) \propto P(X = x, MB(X) = mb(X)).$$

Si denoti con $\mathcal{X}$ l'insieme di tutte le possibili configurazioni delle variabili nella rete. Si consideri inoltre $\mathcal{X}' \subset \mathcal{X}$ sottoinsieme di quelle configurazioni $\mathbf{x}$ in cui $X = x$ e $MB(X) = mb(X)$. Più precisamente:

$$P(X = x, MB(X) = mb(X)) = \sum_{\mathbf{x} \in \mathcal{X}'} P(\mathbf{x})$$

Sia $\phi_c(\mathbf{x_c})$ il fattore della distribuzione rispetto al sottoinsieme $\mathbf{x_c}$

$$\begin{aligned} P(X = x, MB(X) = mb(X)) &= \sum_{\mathbf{x} \in \mathcal{X}'} \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{x_c}) \\ &= \frac{1}{Z} \sum_{\mathbf{x} \in \mathcal{X}'} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{x_c}) \\ &\propto \sum_{\mathbf{x} \in \mathcal{X}'} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{x_c}) \end{aligned} \tag{3.1}$$

dove l'ultimo passo è giustificato dal fatto che Z è costante per un dato *Markov Random Field*. Siccome $\prod_{c \in \mathcal{C}} \phi_c(\mathbf{x_c})$ è costante per ogni $\mathbf{x} \in \mathcal{X}'$ l'ultima formula può essere riscritta come segue:

$$\sum_{\mathbf{x} \in \mathcal{X}'} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{x_c}) = \prod_{c \in \mathcal{C}_X} \phi_c(\mathbf{x_c}) \sum_{\mathbf{x} \in \mathcal{X}'} \prod_{c' \notin \mathcal{C}_X} \phi_{c'}(\mathbf{x_{c'}})$$

quindi si ha che

$$P(X = x|mb(X)) \propto \prod_{c \in \mathcal{C}_X} \phi_c(\mathbf{x_c}) \sum_{\mathbf{x} \in \mathcal{X}'} \prod_{c' \notin \mathcal{C}_X} \phi_{c'}(\mathbf{x_{c'}}).$$

Dato che i sottoinsieme di nodi che non sono in $\mathcal{C}_X$ non contengono X, segue che $\sum_{\mathbf{x} \in \mathcal{X}'} \prod_{c' \notin \mathcal{C}_X} \phi_{c'}(\mathbf{x_{c'}})$ rimane costante per ogni valore x di X.

L'equazione può essere riscritta come:

$$P(X = x|mb(X)) \propto \prod_{c \in \mathcal{C}_X} \phi_c(\mathbf{x_c}).$$

3.3 Inferenza

I modelli grafici mettono a disposizione la possibilità di calcolare la distribuzione di probabilità congiunta sull'insieme di nodi $X \in \mathcal{X}$; possono essere utilizzati per diverse tipologie di scenari. Una situazione standard si ha quando si vuole calcolare la distribuzione di probabilità di un sottoinsieme di nodi $\mathbf{Y}$ condizionata alla conoscenza dei restanti nodi $\mathbf{E}$, $P(\mathbf{Y}|\mathbf{E} = e)$. In questo caso si ha un sottoinsieme di variabili aleatorie $\mathbf{E}$ di cui si conoscono le istanze e ed un sottoinsieme di *query* $\mathbf{Y}$ di cui si vuole calcolare la probabilità condizionata:

$$P(\mathbf{Y}|\mathbf{E} = e) = \frac{P(\mathbf{Y}, e)}{P(e)}$$

ovvero la distribuzione di probabilità sui valori y di $\mathbf{Y}$ condizionata al fatto che $\mathbf{E} = e$. Un altro tipo di scenario si ha quando si vuole trovare l'assegnamento più probabile ad un determinato sottoinsieme di variabili. Come per il calcolo della probabilità condizionata, anche in questo caso si ha un insieme di evidenza $\mathbf{E}$ di cui si conoscono le istanze e. In questo caso, si vuole calcolare quale siano i valori di y maggiormente probabili conoscendo le istanze e. Questo scenario ha due varianti di interesse di cui una è un caso speciale dell'altra.

Nella prima variante, che si chiama *Most Probable Explanation*, si vuole trovare l'assegnamento più probabile a ciascuna variabile aleatoria di cui non si abbia evidenza ovvero di cui non si conosca l'istanza. Sia $\mathbf{W} = \mathcal{X} - \mathbf{E}$, l'obiettivo è quello di trovare l'assegnamento più probabile alle variabili in $\mathbf{W}$ data l'evidenza $\mathbf{E} = e$, ovvero:

$$argmax_w P(W, E)$$

Nella seconda variante, detta *Maximum a posteriori*, si ha sempre un insieme di evidenza $\mathbf{E} = e$ e un sottoinsieme di variabili $\mathbf{Y}$ che non corrisponde

necessariamente all'insieme delle variabili restanti. L'obiettivo rimane quello di trovare l'assegnamento più probabile alle variabili in $\mathbf{Y}$ data l'evidenza $\mathbf{E} = e$, ovvero:

$$argmax_y P(Y, E).$$

Si nota come questa variante sia la generalizzazione del caso precedente in quanto: sia $\mathbf{Z} = \mathcal{X} - \mathbf{Y} - \mathbf{E}$, nel caso generale l'obiettivo risulta calcolare:

$$argmax_y \sum_{\mathbf{Z}} P(Y, Z|e).$$

In generale i modelli grafici possono essere utilizzati per risolvere tutti questi scenari. In modo molto semplice, per calcolare la distribuzione condizionata nel primo scenario, occorre sommare su tutte le possibili configurazioni. Nel secondo e terzo scenario si ricerca tra tutte le possibili configurazioni l'assegnazione più probabile. Tuttavia questo tipo di inferenza esatta non è computazionalmente trattabile in quanto il numero di configurazioni possibili aumenta in modo esponenziale con l'aumento dei nodi nel modello.

Assumendo che si abbia un insieme di fattori $\mathcal{F}$ su un insieme di variabili $\mathcal{X}$, si può definire una funzione non normalizzata:

$$P_{\mathcal{F}}(\mathcal{X}) = \prod_{\phi \in \mathcal{F}} \phi.$$

Per una rete di Markov $\mathcal{G}$, $P_{\mathcal{F}}$ è la versione non normalizzata della distribuzione prima della divisione per la funzione di partizione. È importante notare che la maggior parte delle operazioni che possono essere effettuate su una distribuzione normalizzata possono essere fatte anche sulla distribuzione non normalizzata. In questo modo si può marginalizzare $P_{\mathcal{F}}$ su un sottoinsieme di variabili facendo la somma su tutte le altre. È anche possibile calcolare

$$P_{\mathcal{F}}(\mathbf{X}|\mathbf{Y}) = \frac{P_{\mathcal{F}}(\mathbf{X}, \mathbf{Y})}{P_{\mathcal{F}}(\mathbf{Y})}$$

in cui si nota come sia possibile calcolare la probabilità condizionata senza utilizzare il fattore di normalizzazione.

Nel caso peggiore la complessità dell'operazione di inferenza è intrattabile. Se si assume che l'insieme dei fattori del modello grafico che definiscono la

distribuzione possa essere specificata con un numero polinomiale di bit si arriva al seguente risultato:

Teorema 3.3.1 *I seguenti problemi decisionali sono NP completi:*

- *Data una distribuzione $P_{\mathcal{F}}$ su $\mathcal{X}$, una variabile $X \in \mathcal{X}$ e una istanza $x \in Val(X)$ decidere se $P_{\mathcal{F}}(X = x) > 0$ è NP completo.*
- *Data una distribuzione $P_{\mathcal{F}}$ su $\mathcal{X}$ e un numero τ, decidere se esiste un assegnamento $\mathbf{x}$ di $\mathcal{X}$ tale che $P_{\mathcal{F}} > 0$ è Np completo*

Il seguente problema è #P completo:

- *Data una distribuzione $P_{\mathcal{F}}$ su $\mathcal{X}$, una variabile $X \in \mathcal{X}$ e una istanza $x \in Val(X)$, calcolare $P_{\mathcal{F}}(X = x)$ è #P completo.*

Questi risultati sembrano essere sconfortanti poiché tutti i tipi di inferenza nei modelli grafici sono problemi NP completi o più difficili. Tuttavia, la crescita esponenziale del peso computazionale dell'inferenza può essere evitata utilizzando metodi di inferenza approssimata. Esistono diversi studi per il calcolo dell'inferenza approssimata che utilizzano algoritmi di *sampling*. Gli algoritmi di *sampling* come *Markov Chain Monte Carlo (MCMC)* o *Gibbs sampling* effettuano il calcolo della distribuzione tramite il calcolo ricorsivo della distribuzione con configurazioni casuali. Questi algoritmi derivati dall'algoritmo di Metropolis vengono definiti in modo da ottenere una approssimazione della distribuzione dopo un numero limitato di iterazioni. Le tecniche di inferenza approssimata esulano dallo scopo di questo lavoro di tesi, in cui ci si è concentrati a valutare le potenzialità delle tecniche basate su modelli grafici allo scopo di raffinare le confidenze dei *concept detector*.

L'utilizzo di tecniche di inferenza approssimata risulta invece necessario nel caso in cui si prenda in analisi un largo numero di concetti e viene proposto come possibile sviluppo futuro del presente lavoro di tesi.

Capitolo 4

Approccio Proposto

In questo capitolo viene descritto in modo analitico l'approccio proposto per il raffinamento delle confidenze dei diversi concetti in un insieme di *shot* video. Le confidenze sono state estratte dai video tramite i *concept detector* creati da Mediamill [25]. I *concept detector* sono stati addestrati su un insieme di video messo a disposizione della comunità scientifica da *American National Institute of Standards and Tecnology* (NIST) chiamato *Text REtrievalConference Video Retrival Evaluation* (TRECVID) [22].

4.1 Raffinamento Temporale

La correlazione temporale si riferisce al fatto che, se in un video viene rilevata la presenza di un concetto, è probabile che questa si ripeta anche negli istanti precedenti e successivi. Tra gli *shot* di uno stesso video esiste spesso una coesione temporale per cui se un concetto è presente ad un istante temporale t allora esiste una probabilità che lo stesso concetto sia presente anche agli istanti temporali vicini.

Questa considerazione va fatta anche in relazione alle dinamiche di ripresa di un video. Capita di frequente che un concetto come una scena o un oggetto, venga inquadrato in diversi *shot* vicini temporalmente tra loro in quanto soggetto inerente alla scena. Tra due diversi *shot*, per definizione, si

ha uno stacco della ripresa della videocamera per cui la scena può cambiare notevolmente. Considerando però la coerenza temporale che è presente nella maggior parte dei contenuti video, e che il numero di *shot* è di norma notevolmente maggiore del numero di cambi scena, si può intuire come concetti presenti ad un determinato *shot* t possano essere presenti anche in uno *shot* $t + k$. Tuttavia se tra due diversi *shot* si ha anche solo un cambio di angolatura le caratteristiche di basso livello possono cambiare notevolmente. Considerando che i *concept detector* allo stato attuale non hanno un buon grado di affidabilità può capitare che, anche per *shot* con la stessa scena si abbiano livelli di confidenza piuttosto diversi. Oppure può capitare che vi siano cambi di inquadratura in cui i concetti inquadrati non variano ma la struttura rilevata dalle caratteristiche di basso livello risulta completamente diversa.

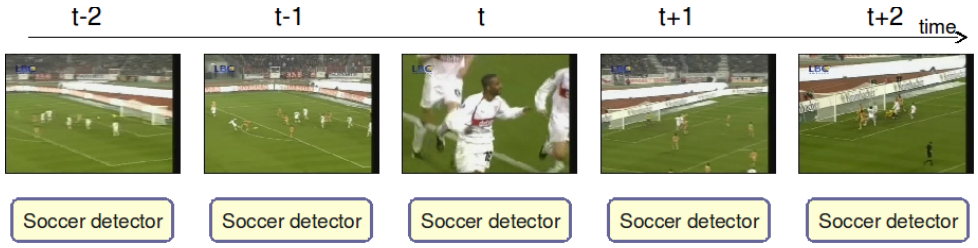

Figura 4.1: Esempio di continuità temporale nonostante cambio di inquadratura

In figura 4.1 si vede come l'inquadratura momentanea del giocatore in primo piano può risultare molto diversa dagli *shot* vicini, tuttavia il concetto *football* rimane costante in tutti gli *shot*.

Per questo motivo risulta conveniente effettuare uno *smooth* sulle confidenze rilevate dai *detector* in modo da smussare le discontinuità che possono essere presenti dall'analisi dei singoli *shot*.

Il raffinamento temporale delle confidenze avviene mediante un approccio che si basa sul lavoro di Liu *et al.* [18]. Ciascuna confidenza viene sottoposta ad una sorta di filtro di *smooth* rispetto alle confidenze in un intorno temporale tramite una combinazione lineare pesata.

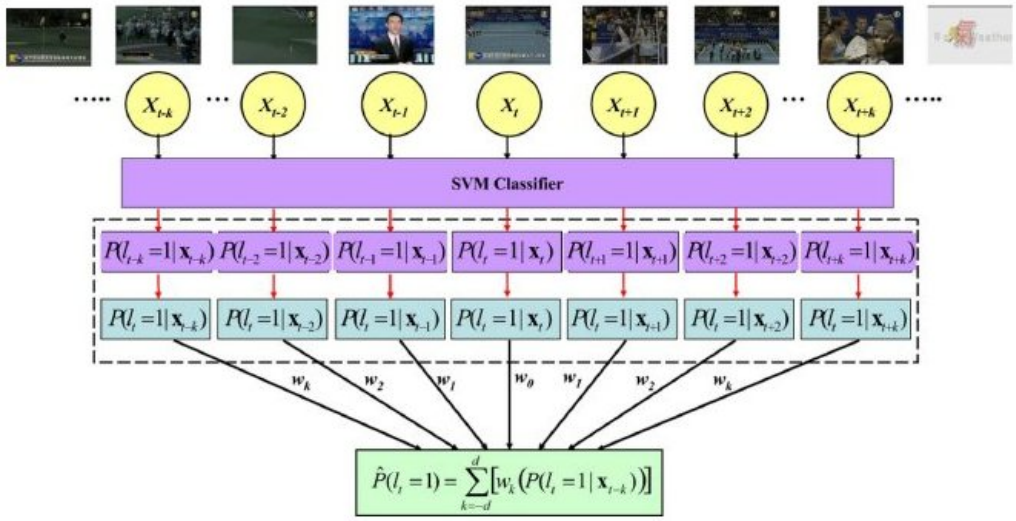

Figura 4.2: Raffinamento temporale tramite combinazione pesata delle confidenze in un intorno temporale

Per ogni concetto si ha a disposizione un insieme di apprendimento da cui è possibile estrarre le informazioni per effettuare il raffinamento. Si hanno a disposizione sia i valori di confidenza che i valori di verità ovvero la corretta valutazione (0 o 1) della presenza del concetto nello *shot*.

Data una finestra temporale di *2D shot*, per ogni concetto, viene stimata statisticamente (sui valori dell'insieme di verità) la probabilità che il concetto sia presente al tempo t dato che il concetto sia presente al tempo t-k e la probabilità che il concetto sia presente al tempo t dato che il concetto sia assente al tempo t-k.

$$P(w_t = 1 | w_{t-k} = 1) = \frac{\#(w_t = 1, w_{t-k} = 1)}{\#(w_{t-k} = 1)}$$

$$P(w_t = 1 | w_{t-k} = 0) = \frac{\#(w_t = 1, w_{t-k} = 0)}{\#(w_{t-k} = 0)}$$

dove w_t è una variabile booleana che indica se nello *shot* t è presente il concetto che si sta analizzando, $\#(w_{t-k} = 1), \#(w_{t-k} = 0)$ sono il numero totale di *shot* rilevanti e irrilevanti nell'insieme di train, $\#(w_t = 1, w_{t-k} = 0)$ è il

numero totale di *shot* in cui si ha che il concetto è rilevante al tempo t mentre lo era rilevante al tempo *t-k* e k appartiene all'insieme *-D,...,-1,1,...,D*. Queste stime vengono utilizzate per migliorare le confidenze nell'insieme di test. Per ogni *shot* si calcola una nuova confidenza della presenza del concetto in analisi come combinazione pesata delle confidenze degli *shot* temporalmente vicini.

$$\begin{aligned}\hat{P}(w_t = 1) = \sum_{k=-d}^{d} \alpha_k P(w_t = 1|f_{t-k}) = \\ \sum_{k=-d}^{d} \alpha_k [P(w_t = 1|w_{t-k} = 1)P(w_{t-k} = 1|f_{t-k}) + \\ P(w_t = 1|w_{t-k} = 0)P(w_{t-k} = 0|f_{t-k})] = \\ \sum_{k=-d}^{d} \alpha_k [P(w_t = 1|w_{t-k} = 1)P(w_{t-k} = 1|f_{t-k}) + \\ P(w_t = 1|w_{t-k} = 0)(1 - P(w_{t-k} = 1|f_{t-k}))]\end{aligned} \tag{4.1}$$

dove f_t indica il valore di confidenza estratto da uno *shot* al tempo t, $P(w_{t-k} = 1|f_{t-k})$ indica la probabilità che il concetto sia presente al tempo *t-k* dato il valore di uscita del detector e α_k è un coefficiente dipendente dal concetto. È necessario stimare ragionevolmente sia la dimensione della finestra temporale sia i coefficienti utilizzati come pesi nella combinazione lineare.

Gli esperimenti effettuati in [18] suggeriscono la finestra temporale D uguale a venti *shot*. Negli esperimenti effettuati nell'applicazione realizzata per questa tesi la finestra D è impostata uguale a 10 poiché le prestazioni con una finestra temporale di dimensioni maggiori risultano peggiori. Questo è dovuto all'utilizzo di una funzione di generazione dei coefficienti diversa da quella utilizzata in [18]. Per quanto riguarda la stima dei coefficienti in [18] viene utilizzata una misura statistica scelta empiricamente. Viene utilizzato un test chi-quadro con livello di confidenza uguale al 99.9% scartando quegli *shot* il cui valore del test chi-quadro è minore di 10,82. Nell'applicazione sviluppata per questa tesi i coefficienti vengono estratti da una funzione

creata ad hoc mostrata in figura 4.3.

$$\alpha_k := \begin{cases} \frac{(1-\alpha_0)2^k}{2-2*(\frac{1}{2})^D}, & \text{per } k \neq 0, \\ 0.5, & \text{per } k = 0. \end{cases}$$

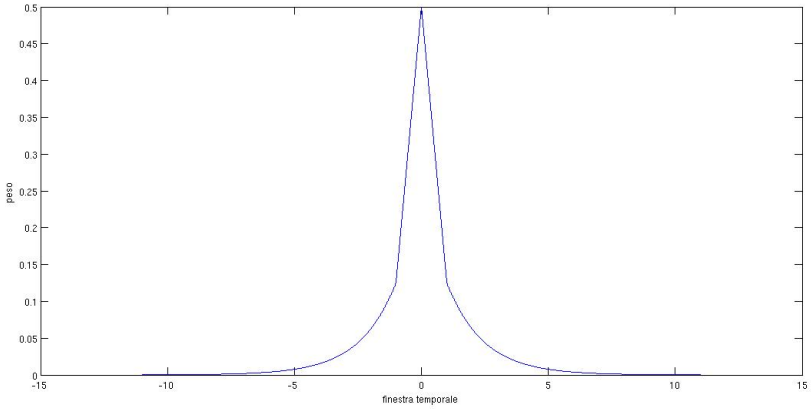

Figura 4.3: Funzione generatrice dei pesi per lo *smooth* temporale

4.2 Raffinamento Semantico

La relazione di occorrenza si riferisce all'informazione che è possibile inferire dalle relazioni che intercorrono tra due diversi concetti nello stesso *shot*. In una immagine presa da uno *shot* video vengono di norma inquadrati diversi concetti. Le relazioni che intercorrono tra coppie di concetti possono essere di diversa natura ma sono riconducibili a due macro-categorie:

- Relazioni semantiche che si riferiscono alla attinenza tra i concetti. Esistono numerose tipologie di relazioni semantiche come:
  - Meronimia (A è parte di B, quindi B ha A come sua parte)
  - Olonimia (B è parte di A, quindi A ha B come sua parte)
  - Iponimia (o troponimia) (A è subordinata a B; A è un tipo di B)

  - Iperonimia (A è subordinata a B)
  - Sinonimia (A denota la stessa cosa di B)
  - Antonimia (A denota il concetto opposto a B)

- Relazioni di osservabilità che si riferiscono alla probabilità statistica di osservare concetti che occorrono nello stesso *shot* anche se non hanno tra loro una relazione semantica. Per esempio il concetto *Montagna* non è semanticamente in relazione con il concetto *Cielo*, tuttavia hanno una relazione di osservabilità poiché, statisticamente, quando in un video è visibile una montagna è possibile osservare anche il cielo.

Considerando entrambe le macro-categorie si nota come le relazioni possono essere sia simmetriche (ovvero che valgono per entrambi i concetti) sia asimmetriche (ovvero che valgono solo per un concetto e non per l'altro). La conoscenza delle relazioni che intercorrono tra i concetti può essere sfruttata per migliorare la corretta percezione da parte del sistema della presenza o assenza dei concetti stessi. Anche il cervello degli esseri umani utilizza le relazioni apprese dal contesto per avere una migliore comprensione delle informazioni ottenute attraverso la vista. Il sistema visivo umano infatti non produce una informazione esaustiva per la comprensione della scena e deve essere effettuato un lavoro di fusione di conoscenza ad opera del cervello. Esso presenta un'enorme e istantanea capacità di collegare diversi concetti ed è in grado di comprendere ciò che stà guardando anche in caso di ambiguità visive.

I difetti di informazione derivati dall'analisi dei video tramite concetti di basso livello devono quindi essere sopperiti tramite un livello di analisi superiore. È necessario che si utilizzi un sistema di apprendimento che comprenda e immagazzini le informazioni a disposizione utilizzandole poi per migliorare le prestazioni di osservazione e di valutazione dei concetti nella scena.

Il raffinamento semantico viene implementato sfruttando la teoria dei modelli grafici, in particolare utilizzando un *Markov Random Field* (o stati di Gibbs, Rete di Markov). Attraverso di questi si vuole realizzare una funzione di fusione di probabilità tra confidenze di concetti diversi estratti

nel medesimo *shot*. Siano M i detector di concetti a disposizione per ogni *shot* si consideri un vettore di verità $\mathbf{w} \in \Omega = \{0,1\}^M$. Per uno *shot* con vettore di verità $\mathbf{w}$ i *detector* assegnano un vettore di valori di verità $f(\boldsymbol{w}) \in [0,1]^M$.

L'obiettivo che si intende realizzare è integrare l'insieme di apprendimento composto da diverse coppie $(\mathbf{w}, f(\boldsymbol{w}))$ e le confidenze $\tilde{f} = f(\tilde{\boldsymbol{w}})$ ottenute dai detector per nuovi *shot*, con vettori di valori di verità incogniti $\tilde{\mathbf{w}}$. Si vuole creare un modello di probabilità dal training set, in particolare un *Markov Random Field* come modello per la distribuzione ricavata dall'insieme di apprendimento.

I *Markov Random Field* si caratterizzano dalla funzione di potenziale che viene utilizzata. In letteratura, sono stati proposti diversi approcci che sfruttano i *Markov Random Field* come mostrato nel capitolo 2.

Un modello che potrebbe essere utilizzato per estrarre le relazioni tra i concetti è il *Markov Random Field* che è descritto da una distribuzione di probabilità che usa come funzione di energia del potenziale il modello di Ising. Il modello di Ising (dal nome del fisico Ernst Ising che lo ha ideato) è un modello fisico-matematico studiato in meccanica statistica. Inizialmente ideato per descrivere un corpo magnetizzato a partire dai suoi costituenti elementari, il modello è stato poi impiegato per modellare fenomeni variegati, accomunati dalla presenza di singoli componenti che, interagendo a coppie, producono effetti collettivi. Il modello di Ising è definito su un insieme discreto di variabili, libere di assumere i valori 1 o -1, che costituiscono i nodi di un reticolo. Possiamo immaginare ciascun nodo come un atomo il cui momento magnetico elementare o *spin* può allinearsi in due direzioni, su (+1) o giù (-1). I nodi interagiscono a coppie: l'energia ha un dato valore quando i due nodi della coppia sono uguali e un altro quando sono diversi. L'energia del reticolo di Ising è definita come:

$$E = \sum_i h_i w_i - \sum_{i,j} J_{ij} w_i w_j$$

dove la somma conta ogni coppia di nodi solo una volta. Notiamo che il prodotto dei nodi è o +1 se i due spin sono uguali (allineati), o -1 se sono diversi (anti-allineati). Il parametro J risulta pari a metà della differenza in

energia tra i due casi. Il *Markov Random Field* che risponde al modello di Ising è descritto dalla seguente distribuzione di probabilità:

$$P_{I,h,J}(w') = \frac{1}{Z} e^{\sum_{i=1}^{M} h_i w'_i + \sum_{i,j=1,\dots,M, i \neq j} J_{i,j} w'_i, w'_j} \tag{4.2}$$

dove h_i possono essere chiamati campi e definiscono i potenziali associati al nodo i, $J_{i,j}$ sono dette interazioni e definiscono i potenziali associati alle relazioni tra i nodi e Z è un fattore di normalizzazione detto funzione di partizione. Per questa distribuzione di probabilità vale la proprietà di Markov dell'indipendenza condizionale per cui la probabilità che $w_i = 1$, data la configurazione del *Markov Random Field* per ogni $j \neq i$, dipende solo dai nodi j adiacenti.

Come visto in sezione 3.2:

$$P(X = x | mb(X)) \propto \prod_{c \in \mathcal{C}_X} \phi_c(\mathbf{x_c})$$

ovvero la probabilità che un nodo X si trovi nello stato x data la conoscenza del rispettivo intorno chiamato *Markov Blanket* è proporzionale al prodotto dei fattori della distribuzione di probabilità di tutte le cricche che contengono X. Nel caso di studio il *Markov Blanket* del concetto i è l'insieme di tutti i restanti concetti meno il concetto i. Si ha quindi:

$$P_{I,h,J}(w'_i = 1 | w'_{M \setminus i}) \propto e^{h_i w'_i + \sum_{j : J_{i,j} \neq 0} J_{i,j} w'_i w'_j}$$

che normalizzata risulta:

$$P_{I,h,J}(w'_i = 1 | w'_{M \setminus i}) = \frac{e^{h_i w'_i + \sum_{j : J_{i,j} \neq 0} J_{i,j} w'_i w'_j}}{e^{h_i + \sum_{j : J_{i,j} \neq 0} J_{i,j} w'_i w'_j} + e^{-h_i - \sum_{j : J_{i,j} \neq 0} J_{i,j} w'_i w'_j}}$$

I campi di Markov possono essere definiti anche con interazioni a più elementi, ovvero con relazioni che rappresentano le relazioni congiunte tra più nodi. Tuttavia se si usa un numero di interazioni maggiore di due si rischia di ottenere una distribuzione con un enorme numero di parametri. Con un numero elevato di concetti si ottiene un sistema non trattabile per cui spesso ci si limita a considerare le interazioni a coppie come nel modello di Ising.

Il modello fin ora descritto non è adatto allo scopo di estrarre le relazioni di occorrenza in quanto si serve di funzioni simmetriche, mentre l'influenza reciproca dei concetti che si vuole modellare non è soltanto simmetrica ma anche asimmetrica. In questo lavoro di tesi si cercano di estrapolare anche le informazioni di relazione di occorrenza incrociata, che sono frequenti in ambito video, ma che non sono ancora state prese in considerazione in modo esaustivo in letteratura.

Nell'approccio proposto si utilizza un modello simile a quello di Ising proposto dal Prof. A. Gandolfi in modo tale da poter caratterizzare le relazioni incrociate tra i concetti. La distribuzione di probabilità su di un vettore **w** è così definita come:

$$P_{h,J}(w) = \frac{1}{Z} e^{\sum_{i=1}^{M} h_i(2w_i-1)+\sum_{i,j=1,\dots,M,i\neq j} J_{i,j}(w_i,w_j)}$$

dove Z è la funzione di partizione:

$$Z = \sum_{w\in\Omega} P_{h,J}(w).$$

Anche in questo caso h_i possono essere chiamati campi e sono i potenziali associati al nodo i mentre $J_{i,j}(w_i, w_j)$ sono dette interazioni e sono i potenziali associati alle relazioni tra i nodi. Si nota che per ogni coppia di concetti i, j vengono definiti quattro parametri che si riferiscono alle diverse relazioni che si possono venire a creare tra i due concetti. In questo modo è possibile caratterizzare sia le interazioni simmetriche che quelle incrociate. Come per il modello di Ising anche questa è una distribuzione definita su un *Markov Random Field* e quindi vale la proprietà di Markov, ovvero che la probabilità condizionata di $w_i = 1$ dato tutto il resto della configurazione dipende solo dai j tali che $J_{i,j} \neq 0$ e vale:

$$P_{h,J}(w_i = 1|w_{M\setminus i}) = \frac{e^{h_i+\sum_{j:J_{i,j}(1,w_j)\neq 0} J_{i,j}(1,w_j)}}{e^{h_i+\sum_{j:J_{i,j}(1,w_j)\neq 0} J_{i,j}(1,w_j)} + e^{-h_i-\sum_{j:J_{i,j}(0,w_j)\neq 0} J_{i,j}(0,w_j)}}$$

Apprendimento del *Markov Random Field*

Esistono diverse tecniche di apprendimento di parametri delle distribuzioni di probabilità relative a *Markov Random Field* basate sulle realizzazioni ovvero,

nel caso in analisi, sui valori di verità. L'approccio sviluppato per questa tesi non utilizza solo le informazioni dei valori di verità per parametrizzare il modello ma anche le confidenze in uscita dai *detector*.

Viene fatta una considerazione fondamentale su cui si basa l'intera struttura del raffinamento semantico.

Osservazione 4.2.1 *Siano* M *i detector di concetti a disposizione per ogni shot,* $\boldsymbol{w} \in \Omega = \{0,1\}^M$ *i vettori di verità e* $\mathrm{f}(\mathbf{w}) \in [0,1]^M$ *le confidenze estratte dai* concept detector. *I valori delle confidenze* $f(w_i)$ *vengono considerati come confidenze condizionate, ovvero se* $w_i = 1$ *allora* $f(w_i)$ *si può considerare come la confidenza condizionata della presenza del concetto i quando il vettore di verità degli altri concetti è quello dato.*

Utilizzando la proprietà di Markov locale si conosce l'espressione analitica della probabilità che il nodo corrispondente alla variabile aleatoria booleana w_i sia uguale a uno data la conoscenza degli stati dei restanti nodi nella rete.

$$P_{h,J}(w_i = 1 | w_{M\setminus i}) = \frac{e^{h_i + \sum_{j:J_{i,j}(1,w_j)\neq 0} J_{i,j}(1,w_j)}}{e^{h_i + \sum_{j:J_{i,j}(1,w_j)\neq 0} J_{i,j}(1,w_j)} + e^{-h_i - \sum_{j:J_{i,j}(0,w_j)\neq 0} J_{i,j}(0,w_j)}}$$

Il *Markov Random Field* può essere quindi modellato estraendo i parametri della distribuzione di probabilità associata, minimizzando la somma dei residui ovvero la seguente distanza sull'insieme di apprendimento:

$$\sum_{w\in\Omega} d^2(P_{h,J}(w), f(w))$$

dove

$$d^2(f, P_{h,J}(w)) = \sum_{i=1}^{M} \alpha_i (f_i - P_{h,J}(w_i = 1 | w_{M\setminus i}))^2.$$

I parametri da stimare sono i campi h_i che si riferiscono al singolo concetto e le interazioni $J_{i,j}(1,1), J_{i,j}(1,0), J_{i,j}(0,1), J_{i,j}(0,0)$ che si riferiscono alle correlazioni tra ciascun concetto. I valori α_i gestiscono l'affidabilità di ciascun *concept detector* e vengono settati con i valori di *Average Precision* calcolati a priori. Si nota come uno dei parametri possa essere considerato

sempre uguale a 1. Questo perchè nella funzione di probabilità condizionata si possono valutare tre parametri rispetto a uno senza perdere generalità. Il sistema risulta trattabile con il numero di concetti a disposizione in quanto per 100 concetti si devono stimare circa 30000 parametri minimizzando una somma su tutte le configurazioni di *train*. Tuttavia il numero di parametri cresce in modo quadratico rispetto al numero di concetti in analisi. Sia N il numero di concetti si devono stimare N campi h_i e $3N^2$ interazioni $J_{i,j}$.

Si definisca il *Markov Random Field* tramite la distribuzione di probabilità sul vettore di verità **w**

$$P_{h,J}(w) = \frac{1}{Z} e^{\sum_{i=1}^{M} h_i(2w_i-1)+\sum_{i,j=1,\dots,M,i\neq j} J_{i,j}(w_i,w_j)}$$

con probabilità condizionata di $w_i \ = 1$

$$P_{h,J}(w_i = 1|w_{M\setminus i}) = \frac{e^{h_i+\sum_{j:J_{i,j}(1,w_j)\neq 0} J_{i,j}(1,w_j)}}{e^{h_i+\sum_{j:J_{i,j}(1,w_j)\neq 0} J_{i,j}(1,w_j)} + e^{-h_i-\sum_{j:J_{i,j}(0,w_j)\neq 0} J_{i,j}(0,w_j)}}$$

Per semplicità di notazione la probabilità condizionata di $w_i = 1$ può essere riscritta come

$$P_{h,J}(w_i = 1|w_{M\setminus i}) = (1 + e^{-2h_i-\sum_{j:J_{i,j}(1,w_j)\neq 0} J_{i,j}(1,w_j)-\sum_{j:J_{i,j}(0,w_j)\neq 0} J_{i,j}(0,w_j)})^{-1}$$

Il *Markov Random Field* viene definito estraendo tutti i parametri della distribuzione di probabilità associata, il metodo che si è utilizzato è la minimizzazione della somma dei residui sui dati di apprendimento:

$$\sum_{w\in\Omega} d^2(P_{h,J}(w), f(w))$$

dove

$$d^2(f, P_{h,J}(w)) = \sum_{i=1}^{M} \alpha_i (f_i - P_{h,J}(w_i = 1|w_{M\setminus i}))^2$$

Sia la funzione da minimizzare:

$$\sum_{n:w_n\in\Omega} \sum_{i=1}^{M} \alpha_i (f_{i,n} - (1 + e^{-2h_i-\sum_{j:J_{i,j}(1,w_j)\neq 0} J_{i,j}(1,w_j)-\sum_{j:J_{i,j}(0,w_j)\neq 0} J_{i,j}(0,w_j)})^{-1})^2)$$

si definiscano per semplicità di notazione

- $\sum J0 = \sum_{j:J_{i,j}(0,w_j)\neq 0} J_{i,j}(0, w_j)$
- $\sum J1 = \sum_{j:J_{i,j}(1,w_j)\neq 0} J_{i,j}(1, w_j)$

la funzione da minimizzare può essere riscritta come

$$\sum_{n:w_n\in\Omega} \sum_{i=1}^{M} \alpha_i (f_{i,n} - (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1})^2)$$

Come primo approccio si è utilizzato un minimizzatore del tipo simplesso *Nelder & Mead* con scarsi risultati ottenendo diversi minimi relativi a seconda dei valori iniziali immessi. Si è valutato di tentare una strada diversa, utilizzando un risolutore L-BFGS ovvero un algoritmo di ottimizzazione quasi-Newtoniano.

I metodi Quasi-Newton costituiscono una classe di metodi per la minimizzazione non vincolata che richiedono soltanto la conoscenza delle derivate prime. Alcuni metodi di questa classe consentono di determinare il minimo di una funzione quadratica definita positiva in un numero finito di iterazioni, in quanto generano direzioni coniugate. La caratteristica più importante dei metodi Quasi-Newton appare risiedere nel fatto che essi forniscono una "approssimazione" del metodo di Newton che conserva (sotto appropriate ipotesi) una rapidità di convergenza superlineare, pur non richiedendo che venga prodotta un'approssimazione consistente della matrice Hessiana. Il metodo BFGS (Broyden, Fletcher, Golfarb, Shanno) è uno dei più noti metodi Quasi-Newton, ed è comunemente ritenuto uno tra i metodi più efficienti di ottimizzazione non vincolata per problemi di dimensione non elevata. L-BFGS è la versione migliorata per problemi ad alta dimensione in cui si ha un limitato uso di memoria. Questo algoritmo di minimizzazione prevede quindi che vengano forniti una funzione obiettivo da minimizzare e il gradiente ovvero è necessario calcolare le derivate prime rispetto ad ogni variabile. Per ogni concetto x le variabili da stimare sono di due tipi:

- il campo h_x
- l'interazione $J_{x,j}$ che a seconda dallo stato del sistema ovvero dalla presenza-assenza dei concetti x e j può essere $J_{i,j}(0, 0)$, $J_{i,j}(1, 0)$ oppure $J_{i,j}(0, 1)$ poiché si considera $J_{i,j}(0, 0)$ sempre uguale a uno.

le derivate rispetto a h_x risultano:

$$\frac{\partial}{\partial h_x} \sum_{n:w_n \in \Omega} \sum_{i=1}^{M} \alpha_i (f_{i,n} - (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1})^2) = \tag{4.3}$$

$$= \sum_{n:w_n \in \Omega} \frac{-4\alpha_i (f_{i,n} - (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1}) * e^{-2h_i - \sum J1 - \sum J0}}{(1 + e^{-2h_i - \sum J1 - \sum J0})^2}. \tag{4.4}$$

Di seguito vengono riportati i passaggi che hanno portato al risultato 4.4. Dalla equazione 4.3 invertendo le sommatorie e la derivata

$$\sum_{n:w_n \in \Omega} \sum_{i=1}^{M} \frac{\partial}{\partial h_x} \alpha_i (f_{i,n} - (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1})^2 = \tag{4.5}$$

si inverte la derivata con α_i

$$\sum_{n:w_n \in \Omega} \sum_{i=1}^{M} \alpha_i \frac{\partial}{\partial h_x} (f_{i,n} - (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1})^2 = \tag{4.6}$$

si deriva rispetto al quadrato

$$\begin{aligned} &\sum_{n:w_n \in \Omega} \sum_{i=1}^{M} 2\alpha_i (f_{i,n} - (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1}) * \\ &\{\frac{\partial}{\partial h_x} (f_{i,n} - (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1})\} = \end{aligned} \tag{4.7}$$

che può essere riscritto

$$\begin{aligned} &\sum_{n:w_n \in \Omega} \sum_{i=1}^{M} 2\alpha_i (f_{i,n} - (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1}) * \\ &\{\frac{\partial}{\partial h_x} f_{i,n} - \frac{\partial}{\partial h_x} (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1}\} = \end{aligned} \tag{4.8}$$

siccome $\frac{\partial}{\partial h_x} f_{i,n} = 0$ si porta fuori il segno meno

$$\begin{aligned} &\sum_{n:w_n \in \Omega} \sum_{i=1}^{M} -2\alpha_i (f_{i,n} - (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1}) * \\ &\frac{\partial}{\partial h_x} (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1} = \end{aligned} \tag{4.9}$$

si deriva rispetto alla potenza -1

$$\sum_{n:w_n\in\Omega}\sum_{i=1}^{M} -2\alpha_i(f_{i,n} - (1+e^{-2h_i-\sum J1-\sum J0})^{-1}) \\ \frac{-1}{(1+e^{-2h_i-\sum J1-\sum J0})^2} * (\frac{\partial}{\partial h_x}1 + \frac{\partial}{\partial h_x}e^{-2h_i-\sum J1-\sum J0}) = \tag{4.10}$$

siccome $\frac{\partial}{\partial h_x}1 = 0$

$$\sum_{n:w_n\in\Omega}\sum_{i=1}^{M} -2\alpha_i(f_{i,n} - (1+e^{-2h_i-\sum J1-\sum J0})^{-1}) \\ \frac{-1}{(1+e^{-2h_i-\sum J1-\sum J0})^2)} * (\frac{\partial}{\partial h_x}e^{-2h_i-\sum J1-\sum J0}) \tag{4.11}$$

si deriva rispetto all'esponenziale

$$\sum_{n:w_n\in\Omega}\sum_{i=1}^{M} -2\alpha_i(f_{i,n} - (1+e^{-2h_i-\sum J1-\sum J0})^{-1})* \\ \frac{-e^{-2h_i-\sum J1-\sum J0}}{(1+e^{-2h_i-\sum J1-\sum J0})^2)} * \frac{\partial}{\partial h_x}(-2h_i - \sum J1 - \sum J0) = \tag{4.12}$$

si considera che nella sommatoria $\sum_{i=1}^{M} \frac{\partial}{\partial h_x}(-2h_i - \sum J1 - \sum J0) = -2$ solo se i = x mentre $\frac{\partial}{\partial h_x}(-2h_i - \sum J1 - \sum J0) = 0 \quad \forall \quad i \neq x$ per questo si può eliminare la sommatoria e riscrivere

$$\sum_{n:w_n\in\Omega} -4\alpha_i(f_{i,n} - (1+e^{-2h_i-\sum J1-\sum J0})^{-1})\frac{2*e^{-2h_i-\sum J1-\sum J0}}{(1+e^{-2h_i-\sum J1-\sum J0})^2} = \\ \sum_{n:w_n\in\Omega} \frac{-4\alpha_i(f_{i,n} - (1+e^{-2h_i-\sum J1-\sum J0})^{-1}) * e^{-2h_i-\sum J1-\sum J0}}{(1+e^{-2h_i-\sum J1-\sum J0})^2} \tag{4.13}$$

che dimostra il risultato 4.4.

Le derivate rispetto a $J_{x,j}$ risultano:

$$\frac{\partial}{\partial J_{x,j}} \sum_{n:w_n\in\Omega} (\sum_{i=1}^{M} \alpha_i(f_{i,n} - (1+e^{-2h_i-\sum J1-\sum J0})^{-1})^2)^{1/2} = \tag{4.14}$$

$$\sum_{n:w_n\in\Omega} \frac{-2[\alpha_i(f_{i,n} - (1+e^{-2h_i-\sum J1-\sum J0})^{-1})e^{-2h_i-\sum J1-\sum J0}]}{(1+e^{-2h_i-\sum J1-\sum J0})^2]}. \tag{4.15}$$

Di seguito vengono riportati i passaggi che hanno portato al risultato 4.15.

La dimostrazione del risultato 4.15 può essere effettuata con gli stessi passaggi della dimostrazione precedente da 4.3 fino a 4.12, quindi la derivata risulta

$$\frac{\partial}{\partial J_{x,j}} \sum_{n:w_n \in \Omega} (\sum_{i=1}^{M} \alpha_i (f_{i,n} - (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1})^2)^{1/2} = \tag{4.16}$$

$$\begin{aligned} &\sum_{n:w_n \in \Omega} \sum_{i=1}^{M} -2\alpha_i (f_{i,n} - (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1}) * \\ &\frac{-e^{-2h_i - \sum J1 - \sum J0}}{(1 + e^{-2h_i - \sum J1 - \sum J0})^2)} * \frac{\partial}{\partial J_{x,j}} (-2h_i - \sum J1 - \sum J0) = \end{aligned} \tag{4.17}$$

anche in questo caso solo i componenti della sommatoria con $i = x$ sono diversi da zero, in particolare $\frac{\partial}{\partial J_{x,j}}(-2h_i - \sum J1 - \sum J0) = -1$ quindi le derivate rispetto a $J_{x,j}$ risultano

$$\sum_{n:w_n \in \Omega} \frac{-2[\alpha_i (f_{i,n} - (1 + e^{-2h_i - \sum J1 - \sum J0})^{-1}) e^{-2h_i - \sum J1 - \sum J0}]}{(1 + e^{-2h_i - \sum J1 - \sum J0})^2]}. \tag{4.18}$$

4.3 Inferenza

Siano $\hat{h}_i, \hat{J}_{i,j}(1,1), \hat{J}_{i,j}(1,0), \hat{J}_{i,j}(0,1)$ i parametri estratti dall'apprendimento si ha a disposizione un modello grafico ottenuto minimizzando la distanza tra le confidenze dei detector su un insieme di apprendimento. Si vuole utilizzare il *Markov Random Field* per fondere le confidenze di test, per cui non si conoscono i valori di verità. L'idea è quella di associare un'opportuna misura di probabilità $P(w)$ ad ogni configurazione $w \in \Omega = \{0,1\}^M$ del sistema. Tipicamente si tende a spiegare l'approccio immaginando di possedere un *ensamble* e cioè un numero enorme di copie di un sistema fisico che evolvono indipendentemente l'una dall'altra. Quando tutte le copie sono all'equilibrio possiamo pensare $P(w)$ come la probabilità di "estrarre" dall'*ensamble* un sistema che si trova nella configurazione w. Lo spazio Ω è suddiviso in configurazioni uguali w, che descrivono lo stato di ciascun concetto.

È possibile stimare la probabilità di una configurazione **w** per un vettore di confidenze di test $\tilde{f}$ calcolando:

$$P(w) = \frac{e^{-d^2(\tilde{f}, P_{h,J}(w))}}{\sum_{w' \in \Omega}(e^{-d^2(\tilde{f}, P_{h,J}(w))})}$$

Si nota come questa probabilità sia significativa solo se la distanza tra il valore di confidenza di test e il valore ottenuto per la configurazione w dal modello sia bassa. Si può considerare la distanza tra il valore di confidenza e il sistema come una misura di entropia del sistema in riferimento al valore di confidenza di test.

Per stimare quale sia la probabilità di presenza di un concetto w_i, per la teoria delle larghe deviazioni, basta sommare i valori di probabilità elementare delle configurazioni in cui $w_i = 1$.

La probabilità del concetto i nello *shot* si ottiene quindi da:

$$P(w_i) = \sum_{w \in \Omega : w_i = 1} P(w)$$

L'approccio di inferenza esatto prevede l'analisi di tutte le configurazioni di Ω, ovvero 2^N. In questa tesi sono state effettuate prove con sottoinsiemi di undici concetti per cui è possibile calcolare l'inferenza esatta. È necessario tuttavia effettuare inferenza approssimata con algoritmi di tipo Monte Carlo o Gibbs Sampling come descritto in Gong *et al.* [8] se si vuole aumentare il numero di concetti utilizzati.

4.4 Calcolo delle nuove confidenze

Il calcolo delle confidenze finali avviene effettuando in serie il raffinamento temporale e il raffinamento semantico.

Come si vede dalla figura 4.4 viene prima effettuato lo *smooth* temporale e successivamente, sulle confidenze in uscita, viene appresa la distribuzione di probabilità che caratterizza il *Markov Random Field*. Il modello restituisce un valore che rappresenta la probalità che ciascun concetto sia presente e in condizioni ottimali potrebbe essere utilizzato come nuova confidenza.

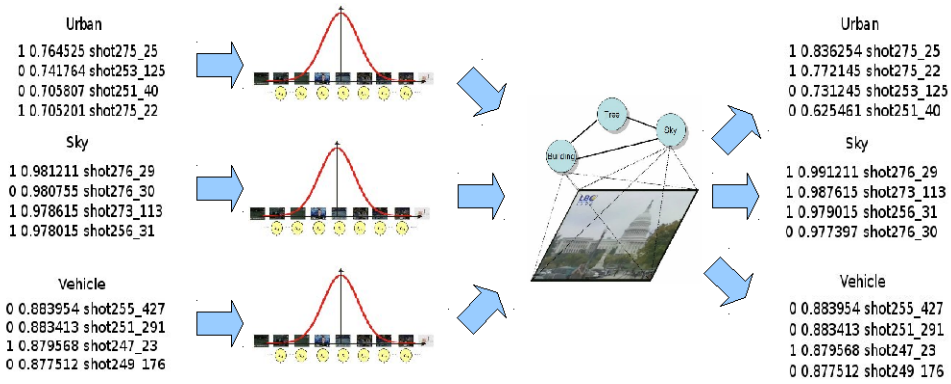

Figura 4.4: Schema di elaborazione delle confidenze

Purtroppo, i *concept detector* contengono una grande quantità di imprecisioni che nell'apprendimento dei parametri vanno a sommarsi restituendo un risultato finale con errori grossolani. L'approccio di minimizzazione della distanza tra i valori di confidenza e il modello non risulta sufficientemente robusto restituendo valori talvolta non attendibili.

Viene quindi utilizzata la previsione del modello in modo tale da sfruttare la capacità di apprendimento delle relazioni senza pesare in modo invasivo sui dati di test.

Per valutare la possibilità di sfruttare le capacità di migliorare le confidenze da parte del sistema, i valori di probabilità restituiti dal modello vengono utilizzati come termini di moltiplicazione con i valori delle confidenze dopo il raffinamento temporale. In una logica di *ranking*, preso un concetto, i valori di confidenza di ciascuno *shot* vengono ordinati per rilevanza e quindi confrontati su tutto l'insieme di test. Sia la probabilità estratta dalla inferenza del *Markov Random Field* $P(w_{i,s}) \in (0,1)$ per il concetto i per uno *shot* s. Si può valutare, considerando tutti i valori di confidenza nell'insieme di test per il concetto i, che esiste un valor medio $E(P(w_i))$. Se il modello di probabilità funziona in modo corretto, sarebbe possibile stimare la presenza/assenza di un concetto ponendo una soglia sul valor medio e annotando come presente gli *shot* che hanno valori di confidenza al di sopra del valor medio e come assente i rimanenti. È possibile quindi valutare il grado di scostamento $\epsilon_{i,s}$

della probabilità in uscita dal modello per ciascuno *shot*, dal valor medio del concetto:

$$P(w_{i,s}) := \begin{cases} E(P(w_i)) + \epsilon_{i,s}, & \text{se il concetto è presente} \\ E(P(w_i)) - \epsilon_{i,s}, & \text{se il concetto è assente.} \end{cases}$$

Se si moltiplica ciascun valore di probabilità in uscita dal modello per il corrispondente valore di confidenza $\tilde{f}_{i,s}$ si ha:

$$\tilde{f}_{i,s} * P(w_{i,s}) = \tilde{f}_{i,s} * (E(P(w_i)) \pm \epsilon_{i,s}) = \tilde{f}_{i,s} * E(P(w_i)) \pm \tilde{f}_{i,s} * \epsilon_{i,s}$$

Si deduce che, pur abbassando le confidenze di un ordine di grandezza, in una logica di annotazione in cui l'accuratezza viene calcolata sulla base della rilevanza dei video in ordine di confidenza, questo approccio migliora le confidenze se il modello restituisce valori corretti. In una logica di *ranking* ciascun concetto viene aumentato o diminuito in proporzione allo scostamento dal valor medio dei valori di probabilità forniti dal modello. Come viene mostrato nel capitolo 5, questo approccio fornisce risultati positivi anche se talvolta non apprezzabilmente ampi. Questo perchè, se le probabilità in uscita dal modello non hanno un sufficiente scarto dal valor medio, le variazioni relative che subiscono i valori di confidenza non sono abbastanza rilevanti da migliorare in modo apprezzabile le prestazioni del sistema.

4.5 Variante

È stata studiata anche una variante al modello di raffinamento relazionale per rendere più robusto e nello stesso tempo più accurato il calcolo delle nuove confidenze. Questa variante viene proposta come possibile sviluppo in quanto in fase di studio prevede la possibilità di dare un contributo di miglioramento mentre in fase realizzativa, in prima analisi, non ha dato risultati che ne giustificassero l'utilizzo.

Nella struttura di probabilità utilizzata nel lavoro di tesi si modella una distribuzione di probabilità a partire dai valori delle confidenze in uscita dai *concept detector* alla presenza/assenza degli altri concetti nella rete. In questa struttura non si considera direttamente la probabilità di presenza/assenza

dei concetti in quanto vengono implicitamente considerati tutti equivalenti e quindi indipendenti. Si vuole introdurre una distribuzione di probabilità anche sulla presenza dei concetti; questa può essere determinata dallo stesso insieme di apprendimento da cui vengono estratti i valori di confidenza ma anche da un universo di immagini e video più ampio. Questo è un aspetto molto interessante in quanto l'annotazione di concetti in immagini e video è largamente a disposizione in archivi digitali per contenuti multimediali in Internet.

La distribuzione di probabilità sulla presenza dei concetti deve essere determinata senza tener conto delle confidenze dei *concept detector*. Un modello ragionevole è la distribuzione di Gibbs che riesce a cogliere anche le interazioni tra i concetti. In modo analogo a quanto descritto in sezione 4.2 è possibile definire una struttura di probabilità su M concetti $\mathbf{w} \in \Omega = \{0,1\}^M$. Si definisce la distribuzione

$$Q_{h,J}(w) = \frac{1}{Z} e^{\sum_{i=1}^{M} h_i(2w_i-1)+\sum_{i,j=1,\ldots,M,i\neq j} J_{i,j}(w_i,w_j)}$$

dove, anche in questo caso, h_i possono essere chiamati campi e definiscono i potenziali associati al nodo i, $J_{i,j}$ sono dette interazioni e definiscono i potenziali associati alle relazioni tra i nodi e Z è un fattore di normalizzazione detto funzione di partizione.

Per le proprietà di indipendenza condizionale di Markov la probabilità condizionata del MRF risulta:

$$Q_{h,J}(w_i = 1|w_{M\setminus i}) = \frac{e^{h_i+\sum_{j:J_{i,j}(1,w_j\neq 0} J_{i,j}(1,w_j)}}{e^{h_i+\sum_{j:J_{i,j}(1,w_j)\neq 0} J_{i,j}(1,w_j)} + e^{-h_i+\sum_{j:J_{i,j}(0,w_j)\neq 0} J_{i,j}(0,w_j)}}.$$

Anche in questo caso per valutare i parametri migliori per caratterizzare il *Markov Random Field* è necessario effettuare una ottimizzazione. L'approccio utilizzato è chiamato *pseudo log-likelihood* e prevede la massimizzazione della probabilità dei dati di apprendimento, dati i parametri che rappresentano il *Markov Random Field*. Sia D una collezione di N *shot* video in cui ciascuno *shot* è caratterizzato da un vettore di verità w indipendente dagli altri *shot*.

$$
\begin{aligned}
-\log(P(D|h,J)) &= -\log(\prod_{s=1}^{N} P(w|h,J)) \\
&= -\log \prod_{s=1}^{N}\prod_{i=1}^{M} Q_{h,J}(w_i = 1|w_{M\setminus i}) = \\
&= -\sum_{s=1}^{N}\sum_{i=1}^{M} \log(Q_{h,J}(w_i = 1|w_{M\setminus i}) = \qquad (4.19)
\end{aligned}
$$

per la proprietà dei logaritmi

sostituendo la probabilità condizionata si ottiene la funzione da minimizzare

$$
-\sum_{s=1}^{N}\sum_{i=1}^{M} \log(\frac{e^{h_i+\sum_{j:J_{i,j}(1,w_j\neq 0} J_{i,j}(1,w_j)}}{e^{h_i+\sum_{j:J_{i,j}(1,w_j)\neq 0} J_{i,j}(1,w_j)} + e^{-h_i+\sum_{j:J_{i,j}(0,w_j)\neq 0} J_{i,j}(0,w_j)}}).
$$

L'ottimizzazione della funzione permette di stimare i parametri $\hat{h}_i$, $\hat{J}_{i,j}(0,0)$, $\hat{J}_{i,j}(1,0)$, $\hat{J}_{i,j}(0,1)$ e $\hat{J}_{i,j}(1,1)$. Sia il modello definito sui valori delle confidenze come descritto in sezione 4.2

$$
P_{h,J}(w) = \frac{1}{Z} e^{\sum_{i=1}^{M} h_i(2w_i-1)+\sum_{i,j=1,\ldots,M,i\neq j} J_{i,j}(w_i,w_j)}
$$

ottenuto minimizzando la seguente somma sul *training set*

$$
\sum_{w\in\Omega} d^2(P_{h,J}(w), f(w))
$$

dove

$$
d^2(f, P_{h,J}(w)) = \sum_{i=1}^{M} \alpha_i (f_i - P_{h,J}(w_i = 1|w_{M\setminus i}))^2
$$

Si può integrare i due modelli assegnando una probabilità che consideri le distanze pesate dal modello sui concetti.

$$
P(w) = \frac{e^{-d^2(\tilde{f}, P_{h,J}(w))} * Q_{h,J}(w)}{\sum_{w'\in\Omega}(e^{-d^2(\tilde{f}, P_{h,J}(w'))} Q_{h,J}(w))}
$$

che espandendo risulta

$$P(w) = \frac{e^{-d^2(\tilde{f},P_{h,J}(w))+\sum_i \hat{h_i}(2w_i-1)+\sum_i \hat{J}_{i,j}(w_i,w_j)}}{\sum_{w'\in\Omega}\left(e^{-d^2(\tilde{f},P_{h,J}(w'))+\sum_i \hat{h_i}(2w_i-1)+\sum_i \hat{J}_{i,j}(w_i,w_j)}\right)}$$

La probabilità del concetto i nello *shot* si otterrebbe quindi da inferenza

$$f(\tilde{w_i}) = \sum_{w\in\Omega:w_i=1} P(w)$$

Si può dare un'interpretazione Bayesiana del modello descritto, a priori si ha una distribuzione di Gibbs dei concetti nelle immagini mentre, a posteriori si ha una valutazione della distribuzione rispetto alle confidenze. In questo senso il fattore $e^{-d^2(\tilde{f},P_{h,J}(w))}$ rappresenta la probabilità condizionata dell'osservazione del valore di confidenza $\tilde{f}$ data la configurazione w. Quindi, a priori si ha una distribuzione di Gibbs che definisce la struttura di probabilità del *Markov Random Field* solo sui concetti, e data una configurazione di presenze di concetti w, i *concept detector* danno un valore che può essere approssimato a $P_{h,J}(w)$.

Capitolo 5

Risultati

Per valutare le performance dell'approccio proposto sono stati effettuati test utilizzando un *dataset* standard largamente utilizzato in ambito di ricerca nell'indicizzazione di contenuti multimediali chiamato TRECVID 2005 [19]. TRECVID è una manifestazione annuale organizzata da *National Institute of Standards and Tecnoligy (NIST)* per promuovere lo sviluppo di tecnologie per il recupero di video digitali tramite analisi di contenuti. La comunità scientifica ha accolto questa iniziativa rendendo di fatto TRECVID il *dataset* standard su cui confrontare i relativi lavori di ricerca. L'insieme di addestramento di TRECVID 2005 consiste di 85 ore di video relativi a notiziari di televisioni arabe, cinesi e americane. Parte di questo insieme viene utilizzato per l'estrazione delle confidenze dai *concept detector* come descritto in *MediaMill* di Snoek *et al.* [25]. L'insieme di apprendimento è suddiviso in due parti in cui il 70% degli *shot* viene utilizzato per l'addestramento dei *concept detector* mentre il restante 30% viene utilizzato per i test, ovvero per determinare i valori di confidenza per ciascun singolo concetto. Questo insieme, che comprende 12.914 *shot*, viene utilizzato in questo lavoro di tesi come insieme di partenza su cui apprendere i modelli ed effettuare i nostri test. Si ha quindi a disposizione un insieme composto da 12.914 *shot*, ognuno dei quali etichettato per ogni concetto con il valore di verità e il valore di uscita del rispettivo *concept detector*.

Per ottenere una valutazione affidabile l'insieme di partenza viene suddi-

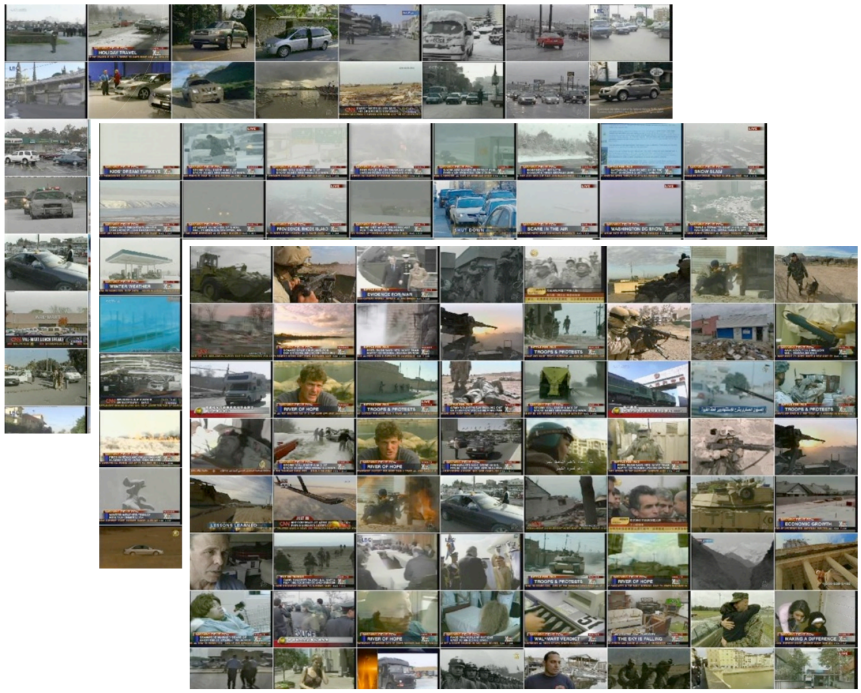

Figura 5.1: Alcuni fotogrammi di esempio dal dataSet TRECVID 2005

viso in modo da effettuare una cross-validazione. La cross-validazione è una tecnica statistica utilizzabile in presenza di una buona numerosità del *training set*. Consiste nella suddivisione del *dataset* totale in k parti (si chiama anche k-fold validation) e, ad ogni passo, la parte (1/k)-esima del *dataset* viene ad essere il validation dataset, mentre la restante parte costituisce il set di apprendimento. Così, per ognuna delle k parti si allena il modello, evitando quindi problemi di overfitting, ma anche di campionamento asimmetrico (e quindi affetto da *bias*) del *training set*, tipico della suddivisione del dataset in due sole parti (ovvero training e validation dataset). In questo lavoro di tesi viene effettuata una *4-fold validation* utilizzando quindi quattro diverse suddivisioni dell'insieme di partenza.

In particolar modo abbiamo scelto di mantenere contigui gli *shot* appartenenti allo stesso video in modo da conservare la relazione temporale preesistente; la suddivisione avviene perciò selezionando un video ogni quattro

e posizionando gli *shot* ad esso relativi nel test set avendo cura di effettuare una rotazione nella selezione in modo da ottenere quattro set diversi nel contenuto.

I valori di uscita dei *concept detector* di *MediaMill* rappresentano la *baseline*, ovvero i valori di riferimento su cui valutare gli eventuali miglioramenti apportati dal nostro contributo.

La precisione del sistema viene calcolata tramite la *Average Precision* (AP) adottata da NIST per misurare, data una *query*, l'accuratezza di risultati di ricerca di concetti, ordinati per pertinenza. Si consideri quindi un insieme di *shot* di test; per ogni concetto viene ordinato (in ordine decrescente) l'insieme in base ai valori delle confidenze. Per valutare la precisione si analizzano i valori di verità di ciascuno *shot* indicando se il concetto è presente (e quindi rilevante) o non è presente.

Si definisce quindi la *Average Precision* come:

$$AP = \frac{1}{min(R,k)} \sum_{j=1}^{k} \frac{R_j}{j} \tilde{w}_j$$

dove R rappresenta il numero dei *shot* rilevanti, R_j il numero di record rilevanti presenti nei primi j record recuperati e $\tilde{w}_j$ è una variabile che assume il valore 1 nel caso il j-esimo record sia rilevante, 0 altrimenti. La variabile k rappresenta un modo per interrompere il processo dopo un preciso numero di record analizzati.

Viene valutata anche la *Mean Average Precision* (MAP) ovvero la precisione media del sistema. Si calcola effettuando la media dei valori di AP rispetto a tutti i concetti analizzati. Questo valore generalizza la precisione del sistema su tutti i concetti e rappresenta l'indicatore "sintetico" delle performance di ricerca del sistema analizzato.

I test per valutare le prestazioni del sistema sono stati effettuati su sottoinsiemi di dieci e undici concetti. Tali sottoinsiemi, mostrati in tabella 5.1, sono stati definiti nel seguente modo:

- Un insieme supervisionato ovvero con concetti scelti appositamente data una logica di interrelazioni. Questo insieme contiene concetti inerenti alla natura che possono co-occorrere in sequenze video.

Tabella 5.1: Concetti utilizzati nei diversi insiemi

Supervisionato	Best_AP	Random1	Random2	Random3	Random4
bird	anchor	tower	monologue	animal	mountain
cloud	face	tree	motorbike	beach	office
grass	fish	truck	mountain	bird	outdoor
house	indoor	urban	natural_disaster	boat	overlayed_text
mountain	newspaper	vegetation	newspaper	cartoon	tower
outdoor	outdoor	vehicle	office	dog	tree
sky	overlayed_text	violence	outdoor	drawing	truck
snow	people	walking_running	overlayed_text	drawing_cartoon	urban
tree	splitscreen	waterbody	people	fish	vegetation
vegetation	studio	waterfall	people_marching	flag	vehicle
weather		weather	police_security	flag_usa	violence

- Un insieme con *baseline* alta per evidenziare le prestazioni su valori di confidenze che hanno una buona *performance* a priori.
- Quattro insiemi casuali per mostrare come il sistema si comporta su concetti relativamente correlati.

Nelle tabelle 5.2,5.3 e 5.4 vengono mostrati i valori di *Average Precision (AP)* ottenute nelle diverse elaborazioni ovvero:

- Nella colonna *BASELINE* sono indicati i valori di precisione di partenza calcolati utilizzando le confidenze non elaborate.
- Nella colonna *RELATION* sono indicati i valori di precisione dopo il raffinamento relazionale.
- Nella colonna *TIME* sono indicati i valori di precisione dopo il raffinamento temporale.
- Nella colonna *TIME+RELATION* sono indicati i valori di precisione dopo aver effettuato prima il raffinamento temporale e successivamente il raffinamento relazionale.

Tabella 5.2: Average Precision dell'insieme Random1

RANDOM1	BASELINE	RELATION	TIME	TIME+RELATION
tower	0.043576739	0.050085152	0.043110112	0.047664528
tree	0.106331427	0.110633991	0.089300405	0.093356440
truck	0.046903364	0.048937612	0.049227952	0.054497919
urban	0.205476925	0.206319884	0.213579789	0.214719261
vegetation	0.150606575	0.153079857	0.160499226	0.161789854
vehicle	0.192351447	0.193674489	0.194555881	0.194825865
violence	0.239357992	0.240995278	0.277451557	0.278370169
walking_running	0.296485131	0.296922677	0.317798563	0.318409607
waterbody	0.145636860	0.142242897	0.178568916	0.178457488
waterfall	0.000912463	0.002867615	0.000875702	0.001497068
weather	0.574627556	0.574213817	0.715281900	0.715127313
MAP	0.182024225	0.183633934	0.203659091	**0.205337774**

In questo capitolo verranno mostrati i risultati di tre insiemi mentre i risultati dei restanti insiemi sono esposti in Appendice A.

La tabella 5.2 mostra le *Average Precision* di un insieme di concetti casuali. Le precisioni di *baseline* variano tra di loro anche in modo sensibile in quanto i *concept detector* a cui si riferiscono hanno comportamenti diversi. Le *Mean Average Precision (MAP)* corrispondono al valore di precisione media ottenuto dal sistema per la relativa elaborazione.

I comportamenti delle *Average Precision* della maggior parte dei concetti vedono un piccolo aumento dopo il raffinamento relazionale e un aumento più sostanziale con il raffinamento temporale. Comportamento diverso si nota in *tree* e *waterfall* in cui si ha un aumento più marcato dopo l'elaborazione relazionale rispetto a quella temporale. Non è facile dimostrare le motivazioni di questi comportamenti. Tuttavia si può supporre che l'aumento di *waterfall* sia in relazione con l'aumento di *tree* e di *vegetation*. Per quanto riguarda l'aumento di prestazioni dovuto al raffinamento temporale si nota

Tabella 5.3: Average Precision dell'insieme supervisionato

Supervisionato	BASELINE	RELATION	TIME	TIME+RELATION
bird	0.383538350	0.376569176	0.404190037	0.404736189
cloud	0.115507342	0.117095639	0.120695504	0.130037930
grass	0.064385397	0.064616232	0.060685532	0.061186407
house	0.010430266	0.007509321	0.010496184	0.008173547
mountain	0.162418590	0.167354275	0.172848198	0.173749231
outdoor	0.686753082	0.686012070	0.725602249	0.725442965
sky	0.463732233	0.464265251	0.471145637	0.471942880
snow	0.004857947	0.019298089	0.004706582	0.013652536
tree	0.106331427	0.106579590	0.089300405	0.092024646
vegetation	0.150606575	0.149030590	0.160499226	0.160472175
weather	0.574627556	0.574352209	0.715281900	0.715143591
MAP	0.247562615	0.248425677	0.266859223	**0.268778372**

il comportamento particolare di *weather*. Questo può essere giustificato dal fatto che il *dataset* in esame è composto da video presi da notiziari in cui solitamente si ha una porzione di programma dedicata esclusivamente al meteo. Una tipologia di video che prevede la presenza di un concetto per una porzione di *shot* contigui e con rare presenze occasionali beneficia in modo ottimale del raffinamento temporale.

Il grafico mostrato in figura 5.3 corrisponde alle *Average Precision* della tabella 5.2. È possibile notare come il contributo maggiore al miglioramento sia dovuto al raffinamento temporale. Il comportamento medio valutato nell'istogramma *MAP* mostra come la precisione abbia un aumento in ciascuna elaborazione e che il valore migliore si ottiene dopo l'elaborazione che prevede entrambi i raffinamenti.

La tabella 5.3 mostra le *Average Precision* dell' insieme di concetti supervisionato. Anche in questo caso si nota dalla *Mean Average Precision* che si ottengono miglioramenti in tutte le elaborazioni nonostante, anche in que-

Tabella 5.4: Average Precision dell'insieme best_AP

best_AP	BASELINE	RELATION	TIME	TIME+RELATION
anchor	0.585068869	0.598528431	0.524372207	0.555567677
face	0.890145126	0.889200550	0.896019698	0.897299463
fish	0.189178506	0.182731453	0.231630696	0.226347286
indoor	0.601190619	0.604392805	0.619020225	0.618177452
newspaper	0.212872795	0.204008379	0.350467089	0.350678583
outdoor	0.686753082	0.679258596	0.725602249	0.721768865
overlayed_text	0.664978983	0.661381161	0.677634564	0.676979157
people	0.841409358	0.883397839	0.855241083	0.861179228
splitscreen	0.549777623	0.583988831	0.496733035	0.506448748
studio	0.644570839	0.648633320	0.661540209	0.662628640
MAP	0.586594580	0.593552136	0.603826105	**0.607707510**

sto insieme, il contributo dato dal raffinamento relazionale sia sensibilmente inferiore. L'aumento di prestazione maggiore per il raffinamento relazionale si ha con il concetto *snow* in cui la precisione di *baseline* è molto bassa. Per quanto riguarda il raffinamento temporale si ha un buon miglioramento oltre che per *weather* anche per *outdoor* e *bird*.

Il grafico mostrato in figura 5.4 corrisponde alle *Average Precision* della tabella 5.3. Si nota come i piccoli miglioramenti in *snow*, *mountain* e *tree* siano corrisposti da piccoli peggioramenti in *vegetation* e *bird* in modo da rendere la *Main Average Precision* praticamente invariata dopo il raffinamento temporale.

La tabella 5.4 mostra le *Average Precision* dell' insieme di concetti con le *baseline* più alte. In questa tabella il raffinamento relazionale ha un comportamento migliore rispetto ai precedenti insiemi. Questo è dovuto al fatto di utilizzare confidenze di *concept detector* che sono più affidabili a priori.

Il grafico mostrato in figura 5.5 corrisponde alle *Average Precision* della tabella 5.4. Si nota come l'istogramma relativo alla *Main Average Precision*

abbia un miglioramento crescente delle prestazioni, ovvero se il raffinamento relazionale apporta un aumento di prestazioni questo si va a sommare a quello dato dal raffinamento temporale.

Urban	Vehicle	Sky
1 0.764525 shot275_25	0 0.883954 shot255_427	1 0.981211 shot276_29
0 0.741764 shot253_125	0 0.883413 shot251_291	0 0.980755 shot276_30
0 0.705807 shot251_40	1 0.879568 shot247_23	1 0.978615 shot273_113
1 0.705201 shot275_22	0 0.877512 shot249_176	1 0.978015 shot256_31

(a) *Urban.* (b) *Vehicle.* (c) *Sky.*

Figura 5.2: Alcune esempi di *shot* con relative confidenze e valori di verità.

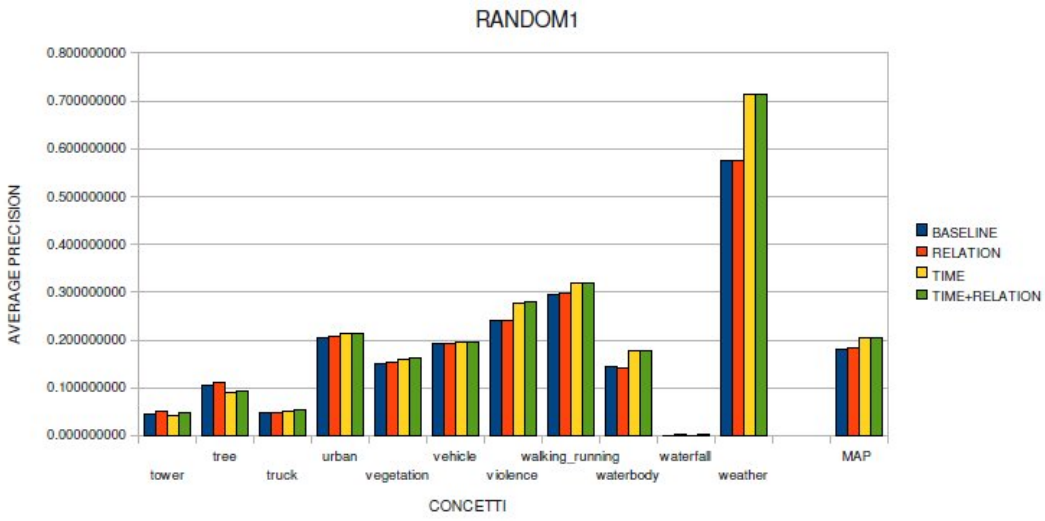

Figura 5.3: Grafico delle *Average Precision* dell'insieme Random 1

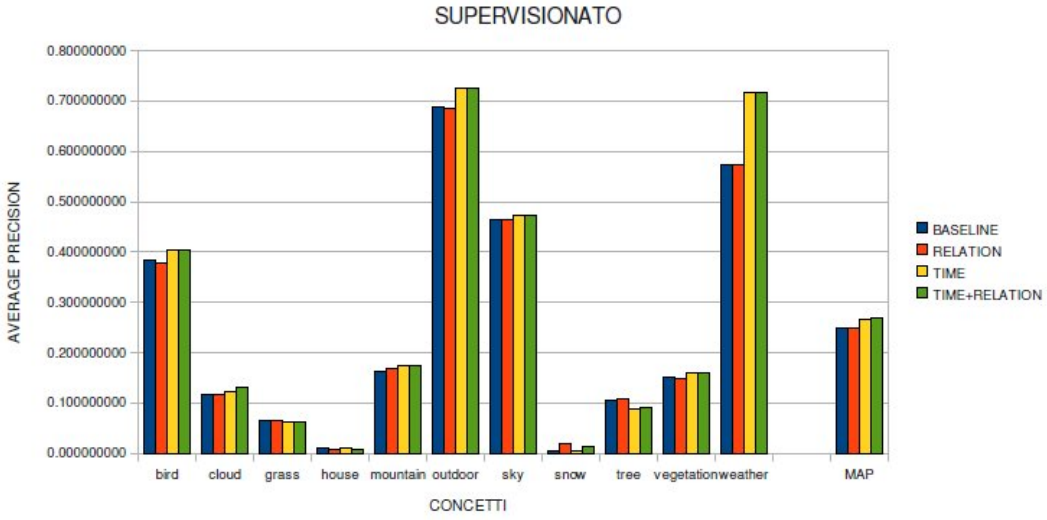

Figura 5.4: Grafico delle *Average Precision* dell'insieme Supervisionato

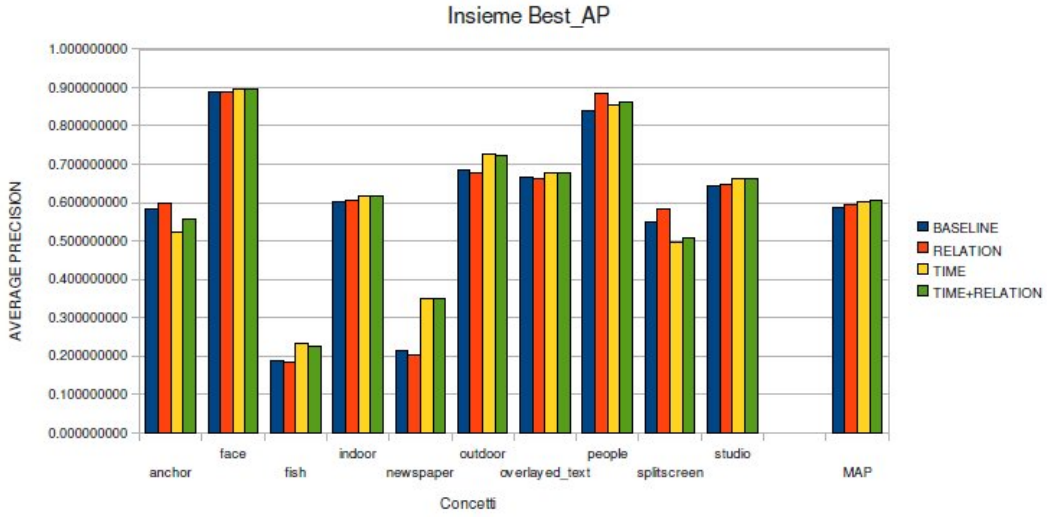

Figura 5.5: Grafico delle *Average Precision* dell'insieme best_AP

Capitolo 6

Conclusioni e Sviluppi futuri

6.1 Conclusioni

In questo lavoro di tesi è stato studiato un approccio per migliorare la precisione di annotazione automatica di video mediante analisi delle co-occorrenze spazio-temporali. È stata proposta e presentata una tecnica volta a migliorare i valori di confidenza di presenza di concetti.

Le confidenze vengono fornite inizialmente dai *concept detector* creati dal progetto *MediaMill* [25] e vengono modificate tenendo in considerazione:

- Coerenza temporale;
- Relazioni tra concetti.

Il raffinamento temporale viene realizzato mediante uno *smooth* delle confidenze. Vengono calcolate in un insieme di apprendimento le probabilità di presenza di un concetto in relazione alla presenza o assenza dello stesso in *shot* temporalmente vicini. Queste probabilità vengono poi utilizzate per ricalcolare ciascun valore di confidenza tramite combinazione lineare delle confidenze negli *shot* temporalmente vicini, pesate per le rispettive probabilità . Le tecniche di raffinamento temporale, tramite combinazione lineare delle confidenze, hanno dimostrato in letteratura di poter apportare effettivi miglioramenti. In particolare, nell'approccio da noi proposto, si ottengono

risultati di miglioramento dell'affidabilità delle confidenze in linea con quelle ottenute in letteratura, aumentando mediamente la precisione di circa il 10%.

Il raffinamento relazionale viene realizzato tramite un *Markov Random Field*, la cui distribuzione di probabilità è stata studiata appositamente per parametrizzare tutte le possibili relazioni che possono intercorrere tra coppie di concetti. Questo approccio, che propone l'idea "ambiziosa" di considerare non solo le co-occorrenze ma tutte le possibili relazioni, è all'avanguardia per quanto riguarda lo stato dell'arte per la fusione di confidenze in ambito multimediale. Per questo motivo può essere considerato uno studio apprezzabile pur non mostrando ancora risultati notevoli, ancorché in linea con lo stato dell'arte in questo tipo di analisi. I problemi riscontrati nel mettere in relazione confidenze provenienti da diversi *concept detector* sono dovuti sia agli errori intrinseci che si moltiplicano nel considerare le relazioni tra le coppie di concetti, sia all'utilizzo di una tecnica di apprendimento dei parametri che non è risultata sufficientemente robusta. Tuttavia, questo lavoro di tesi apre le porte ad uno studio più approfondito per quanto riguarda l'apprendimento robusto dei parametri della funzione di distribuzione di probabilità che caratterizza il *Markov Random Field* in modo da estrarre più rigorosamente le relazioni che intercorrono tra le confidenze. Inoltre, propone lo studio di una variante promettente che mira a caratterizzare sia le relazioni tra le confidenze sia le relazioni tra i concetti, che possono essere utilizzate da qualsiasi fonte di annotazione automatica di concetti nei video.

6.2 Sviluppi futuri

Possibili sviluppi per migliorare il metodo proposto riguardano principalmente il raffinamento relazionale. Come accennato in sezione 4.5 è stata già studiata una modifica sostanziale al modello per migliorare la precisione del sistema. Nella struttura di probabilità utilizzata nel lavoro di tesi si crea un modello a partire dai valori di confidenza in uscita dai *concept detector* alla presenza o meno degli altri concetti nella rete. In questa struttura non si considera direttamente la presenza o assenza dei concetti in quanto vengono

implicitamente considerati tutti equivalenti e quindi indipendenti. Si potrebbe perciò introdurre una distribuzione di probabilità anche sui concetti che potrebbe essere determinata dallo stesso insieme di apprendimento, come da un universo di immagini più ampio. Questo è un aspetto molto interessante in quanto l'annotazione di presenza/assenza di concetti in immagini e video è largamente a disposizione in archivi digitali per contenuti multimediali in Internet.

Un ulteriore sviluppo per migliorare il sistema riguarda il metodo di apprendimento di parametri. In questo lavoro di tesi si è realizzato un sistema di apprendimento per minimizzazione della distanza Euclidea tra i valori di confidenza e la funzione di probabilità condizionata del modello. Sarebbe opportuno utilizzare tecniche di apprendimento più robuste in modo da ottenere parametri più precisi per migliorare l'accuratezza del sistema.

Uno sviluppo che risulta necessario nel caso in cui si voglia utilizzare un insieme più ampio di concetti è l'implementazione di una inferenza approssimata tramite algoritmi del tipo MonteCarlo o *Loopy Belief Propagation*. In questo lavoro di tesi si è implementato una metodo di inferenza esatta che però risulta computazionalmente non trattabile se il numero di concetti è grande. Ciò è dovuto alla necessità di calcolare la funzione di partizione che è una somma su tutte le possibili configurazioni della rete che cresce in modo esponenziale rispetto al numero di concetti.

Appendice A

Appendice

In questo capitolo verranno presentate le tabelle e i grafici delle precisioni degli insiemi di concetti utilizzati come test del sistema non mostrati nel capitolo dei risultati.

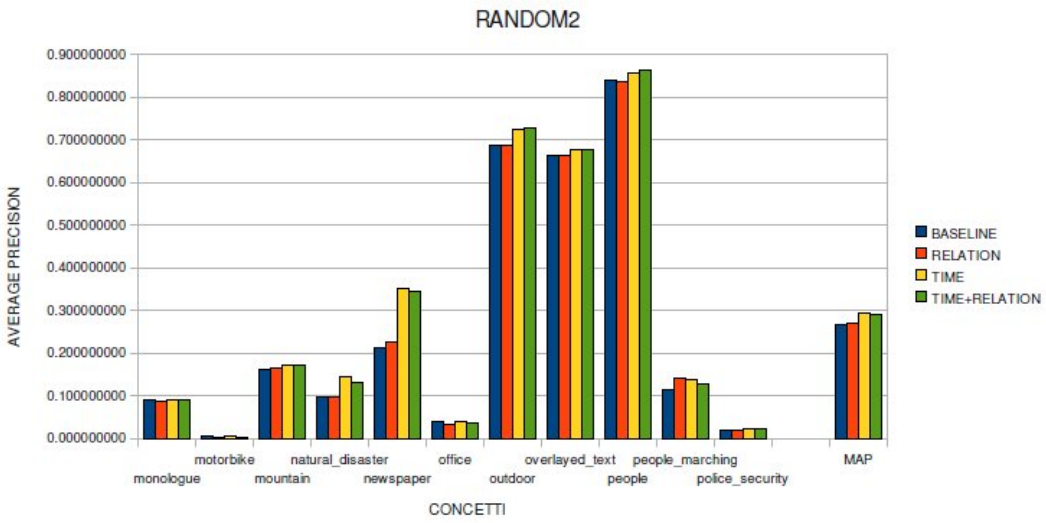

Figura A.1: Grafico delle *Average Precision* dell'insieme Random 2

Tabella A.1: Average Precision dell'insieme Random2

RANDOM2	BASELINE	RELATION	TIME	TIME+RELATION
monologue	0.089274723	0.089094764	0.089304870	0.089514089
motorbike	0.007450338	0.003232705	0.006324236	0.002848251
mountain	0.162418590	0.164911104	0.172848198	0.170677424
natural_disaster	0.098111142	0.095996449	0.145507716	0.130971898
newspaper	0.212872795	0.225809467	0.350467089	0.343216350
office	0.039912191	0.032572129	0.039608411	0.036078480
outdoor	0.686753082	0.687826246	0.725602249	0.726449969
overlayed_text	0.664978983	0.664879516	0.677634564	0.676656971
people	0.841409358	0.835845108	0.855241083	0.865553305
people_marching	0.115703036	0.141580611	0.137664354	0.126130579
police_security	0.019692286	0.019192001	0.024184570	0.022134109
MAP	0.267143320	0.269176373	**0.293126122**	0.290021039

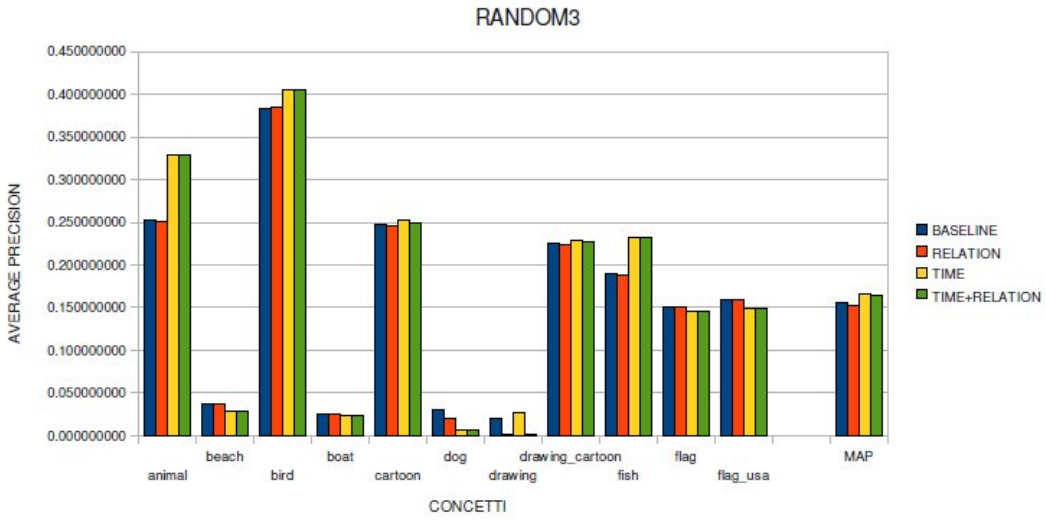

Figura A.2: Grafico delle *Average Precision* dell'insieme Random 3

Tabella A.2: Average Precision dell'insieme Random3

RANDOM3	BASELINE	RELATION	TIME	TIME+RELATION
animal	0.252379769	0.251509682	0.328693498	0.328856257
beach	0.037330188	0.037300557	0.029538026	0.029523566
bird	0.383538350	0.384438360	0.404305938	0.404154625
boat	0.025287946	0.025139013	0.023891765	0.023868882
cartoon	0.247230860	0.245116613	0.252891279	0.249991755
dog	0.030028766	0.020937075	0.006638793	0.007235257
drawing	0.020121739	0.001509706	0.026547796	0.002054073
drawing_cartoon	0.224577416	0.223463562	0.228641686	0.227810858
fish	0.189178506	0.187874536	0.231630696	0.231593242
flag	0.151065251	0.151073806	0.145940344	0.145848950
flag_usa	0.159148218	0.159090958	0.149245126	0.149002658
MAP	0.156353364	0.153404897	**0.166178632**	0.163630920

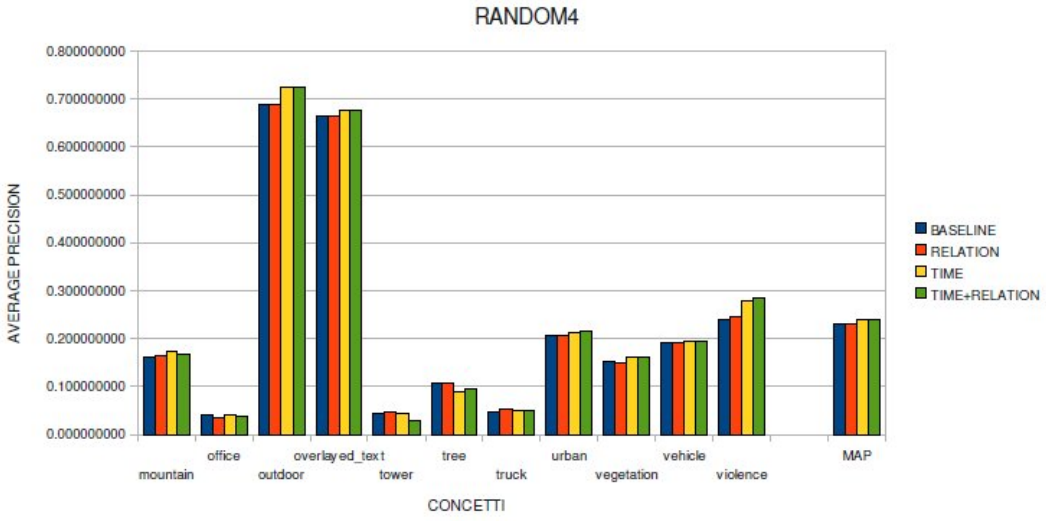

Figura A.3: Grafico delle *Average Precision* dell'insieme Random 4

Tabella A.3: Average Precision dell'insieme Random4

RANDOM4	BASELINE	RELATION	TIME	TIME+RELATION
mountain	0.162418590	0.162679858	0.172848198	0.168580743
office	0.039912191	0.033140127	0.039608411	0.037059339
outdoor	0.686753082	0.687928408	0.725602249	0.726004589
overlayed_text	0.664978983	0.664810748	0.677634564	0.677000903
tower	0.043576739	0.047219749	0.043110112	0.028040988
tree	0.106331427	0.108076896	0.089300405	0.094107073
truck	0.046903364	0.051568368	0.049227952	0.049555324
urban	0.205476925	0.206584215	0.213579789	0.217101319
vegetation	0.150606575	0.149571134	0.160499226	0.161030903
vehicle	0.192351447	0.192862814	0.194555881	0.194553006
violence	0.239357992	0.244465012	0.277451557	0.286051973
MAP	0.230787938	0.231718848	**0.240310758**	0.239916924

Bibliografia

[1] R. B. N. Aly and D. Hiemstra, "Concept detectors: how good is good enough?" in *Proceeding of the 17th ACM International Conference on Multimedia, Beijing, P.R. Republic of China*. New York: ACM, 2009, pp. 233–242.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. Springer, October.

[3] S.-F. Chang, L. S. Kennedy, and E. Zavesky, "Columbia university's semantic video search engine," in *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*. New York, NY, USA: ACM, 2007, pp. 643–643.

[4] M. G. Christel, "Carnegie mellon university traditional informedia digital video retrieval system," in *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*. New York, NY, USA: ACM, 2007, pp. 647–647.

[5] P. Clifford, "Markov random fields in statistics," 1990.

[6] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 1–60, 2008.

[7] C. G.M.Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, J. R. R. Uijlings, M. van Liempt, M. Bugalho, I. Trancoso, F. Yan, M. A. Tahir, K. Mikolajczyk, J. Kittler, M. de Rijke, J.-M. Geusebroek, T. Gevers,

M. Worring, D. C. Koelma, and A. W. M. Smeulders, "The MediaMill TRECVID 2009 semantic video search engine," in *Proceedings of the 7th TRECVID Workshop*, Gaithersburg, USA, November 2009.

[8] Y. Gong and W. Xu, *Machine Learning for Multimedia Content Analysis (Multimedia Systems and Applications)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[9] A. G. Hauptmann, M. yu Chen, M. G. Christel, W.-H. Lin, and J. Y. 0003, "A hybrid approach to improving semantic extraction of news video," in *ICSC*, 2007, pp. 79–86.

[10] Y. G. Jiang, C. W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*. New York, NY, USA: ACM, 2007, pp. 494–501. [Online]. Available: http://dx.doi.org/10.1145/1282280.1282352

[11] L. S. Kennedy and S.-F. Chang, "A reranking approach for context-based concept fusion in video indexing and retrieval," in *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*. New York, NY, USA: ACM, 2007, pp. 333–340.

[12] K. H. Kim, "The theory of matrices : P. lancaster and m. tismenetsky, new york: Academic press, 1985, 570 pages," *Mathematical Social Sciences*, vol. 13, no. 1, pp. 87–87, February 1987. [Online]. Available: http://ideas.repec.org/a/eee/matsoc/v13y1987i1p87-87.html

[13] R. Kindermann, *Markov Random Fields and Their Applications (Contemporary Mathematics ; V. 1)*. American Mathematical Society.

[14] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, August 2009. [Online]. Available: http://is.gd/5ys88

[15] D. Koller, N. Friedman, L. Getoor, and B. Taskar, *Graphical Models in a Nutshell*. MIT Press, 2007. [Online]. Available: http://www.robotics.stanford.edu/ koller/Papers/Koller+al:SRL07.pdf

[16] S. Z. Li, "Markov random field models in computer vision," 1994.

[17] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma, "Dual cross-media relevance model for image annotation," in *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*. New York, NY, USA: ACM, 2007, pp. 605–614.

[18] K.-H. Liu, M.-F. Weng, C.-Y. Tseng, Y.-Y. Chuang, and M.-S. Chen, "Association and temporal rule mining for post-filtering of semantic concept detection in video," *Multimedia, IEEE Transactions on*, vol. 10, no. 2, pp. 240–251, Feb. 2008.

[19] P. Over, T. Ianeva, W. Kraaij, and A. Smeaton, "Trecvid 2005 an overview," in *TREC Video Retrieval Evaluation Online Proceedings*, 2006.

[20] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, 1999, pp. 61–74. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639

[21] G. J. Qi, X. S. Hua, Y. Rui, J. Tang, T. Mei, and H. J. Zhang, "Correlative multi-label video annotation," in *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*. New York, NY, USA: ACM, 2007, pp. 17–26. [Online]. Available: http://dx.doi.org/10.1145/1291233.1291245

[22] A. F. Smeaton, P. Over, and W. Kraaij, "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements," in *Multimedia Content Analysis, Theory and Applications*, A. Divakaran, Ed. Berlin: Springer Verlag, 2009, pp. 151–174.

[23] A. W. M. Smeulders, M. Worring, and S. Santini, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, December 2000. [Online]. Available: http://portal.acm.org/citation.cfm?id=357873

[24] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[25] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*. New York, NY, USA: ACM, 2006, pp. 421–430.

[26] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua, "Bayesian video search reranking," in *MM '08: Proceeding of the 16th ACM international conference on Multimedia*. New York, NY, USA: ACM, 2008, pp. 131–140.

[27] J. Wang, Y. Zhao, X. Wu, and X.-S. Hua, "Transductive multi-label learning for video concept detection," in *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*. New York, NY, USA: ACM, 2008, pp. 298–304.

[28] X.-Y. Wei and C.-W. Ngo, "Fusing semantics, observability, reliability and diversity of concept detectors for video search," in *MM '08: Proceeding of the 16th ACM international conference on Multimedia*. New York, NY, USA: ACM, 2008, pp. 81–90.

[29] M.-F. Weng and Y.-Y. Chuang, "Multi-cue fusion for semantic video indexing," in *MM '08: Proceeding of the 16th ACM international conference on Multimedia*. New York, NY, USA: ACM, 2008, pp. 71–80.

[30] J. Yang and A. Hauptmann, "(un)reliability of video concept detection," in *CIVR '08: Proceedings of the 2008 international conference on*

Content-based image and video retrieval. New York, NY, USA: ACM, 2008, pp. 85–94.

[31] J. Yang and A. G. Hauptmann, "Exploring temporal consistency for video analysis and retrieval." in *Multimedia Information Retrieval*, J. Z. Wang, N. Boujemaa, and Y. Chen, Eds. ACM, 2006, pp. 33–42. [Online]. Available: http://dblp.uni-trier.de/db/conf/mir/mir2006.html

[32] M. yu Chen and A. G. Hauptmann, "Discriminative fields for modeling semantic concepts in video," in *RIAO*, 2007.

[33] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua, "Graph-based semi-supervised learning with multiple labels," *J. Vis. Comun. Image Represent.*, vol. 20, no. 2, pp. 97–103, 2009.

[34] Z.-J. Zha, T. Mei, Z. Wang, and X.-S. Hua, "Building a comprehensive ontology to refine video concept detection," in *MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval*. New York, NY, USA: ACM, 2007, pp. 227–236.

[35] W. Zheng, J. Li, Z. Si, F. Lin, and B. Zhang, "Using high-level semantic features in video retrieval," in *In Proc. of CIVR*. Springer, 2006, pp. 370–379.

[36] Y.-T. Zheng, S.-Y. Neo, T.-S. Chua, and Q. Tian, "Probabilistic optimized ranking for multimedia semantic concept detection via rvm," in *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*. New York, NY, USA: ACM, 2008, pp. 161–168.

Ringraziamenti

I primi ringraziamenti vanno doverosamente al Prof e a Marco, la loro professionalità e umanità sono un esempio da seguire. Un grazie speciale va al Prof. Alberto Gandolfi, senza il quale questo lavoro di tesi non avrebbe parte della sua struttura. Oltre al suo nome così familiare, mi hanno colpito molto la squisita gentilezza e la disponibilità che mi ha dimostrato in questi mesi. Un ringraziamento e un abbraccio a tutti i ragazzi del MICC: il Meoni, il Masi, Beppino, Lorenzo, Fernando, il Dini, il Pernici ma anche il Marto e Filippo... sono orgoglioso di aver fatto parte, anche per poco, di un gruppo di ingegneri davvero d'eccellenza. Un ringraziamento a parte va ai miei correlatori Lambe e Beppone che nelle difficoltà e nei momenti di smarrimento mi hanno sempre sostenuto e incoraggiato.

Vorrei ricordare le persone che mi sono state vicine in questi anni di specialistica, a partire da Marco con cui ho percorso gran parte del cammino universitario e con cui ho condiviso gioie che raramente si vivono durante gli anni universitari. Il Cebe, all'anagrafe Alberto Gandolfi, con lui ho vissuto in questi ultimi anni a Firenze, ma è tutta la vita che ne combiniamo di tutti i colori...e l'Olmo. Un grazie a tutti gli amici e alle amiche che ho conosciuto in questi anni a Firenze in particolare a Sarina, Mimì, alle ragazze della presidenza, Ruggero, Nicolas, Mari e Cosimo, Lara, alle amiche di Parma,a Svetlana e Viola... Un altro grazie a quelli che da Piacenza e dintorni mi hanno sempre aiutato, magari senza saperlo, a portare a termine questo percorso: al Fossa, al Monda, a tutti gli amici di Pontenure che sono troppi da citare ma che sono presenti ovunque vada, alla Silivetta, Cristian, la Fra, la Lory, la Giova e Binaghi e agli amici sparsi in giro per il mondo ma che

ho sempre sentito vicino.

Un abbraccio speciale a $\mathcal{M}$ che mi ha sopportato, aspettato e incoraggiato, che mi ha insegnato a costruire, giorno per giorno, un futuro. Infine un grazie alla mia famiglia che più mi allontano, più sento vicino. In particolare, un ringraziamento ai miei genitori, per la pazienza, la generosità e per avermi sempre dato un appoggio.

Finito di stampare il 7 aprile 2010 utilizzando LaTeX