Università degli Studi di Brescia
Dipartimento di Ingegneria dell'Informazione

Corso di Laurea in Ingegneria Elettronica e delle Telecomunicazioni

RELAZIONE FINALE

Classificazione automatica di inquadrature in film d'autore mediante apprendimento supervisionato

Laureando:
Marco Cominelli Matricola **88971**

Relatore:
Chiar.mo **Prof. Sergio Benini**

Anno Accademico 2015/2016

Compendio

La scelta del tipo di inquadratura rappresenta una delle tecniche cinematografiche principali per sviluppare la narrazione e indurre un coinvolgimento emotivo nello spettatore. Per capire fino a che punto la distribuzione delle inquadrature (Shot Scale Distribution, SSD) possa essere significativa per la rappresentazione di un film, è necessario produrre il riconoscimento automatico delle SSD su un'ampia raccolta di film.

In questo lavoro viene studiato e verificato un possibile approccio per la stima automatica delle SSD. L'analisi dei fotogrammi del film utilizza le proprietà intrinseche dell'immagine per estrapolare informazioni sulla distanza della cinepresa senza la necessità di ricostruire la struttura tridimensionale della scena. La verifica sperimentale delle prestazioni del sistema è stata effettuata su dieci film appartenenti alla filmografia del regista italiano Federico Fellini (1920-1993).

Abstract

The scale of shot is one of the main techniques employed in audiovisual products in order to convey meaning and induce the viewer's emotional state. In order to understand at which level the Shot Scale Distribution (SSD) of a movie might become its fingerprint, it is necessary to produce automatic recognition of shot scale on a large movie corpus.

In this work we study and test an automatic framework for estimating the SSD of a movie by using inherent characteristics of shots containing information about camera distance, without the need to recover the tridimensional structure of the scene. The experimental investigation that shows its validity is conducted on ten movies by Italian director Federico Fellini (1920-1993).

Ringraziamenti

Concludendo questa prima parte del mio percorso universitario, sono veramente tante le persone verso le quali devo i miei più sentiti ringraziamenti.

Prima di tutto intendo ringraziare il Prof. Sergio Benini, relatore di questa tesi, che mi ha guidato nel lavoro mettendo a mia disposizione i mezzi fondamentali per completarlo, e il Dott. Michele Svanera, che mi ha assistito con enorme disponibilità e infinita pazienza.

Voglio poi ringraziare con il cuore tutti gli amici e le persone che mi sono state vicino in questi anni, in particolare Andrea, Cristian, Diego, Marcello, Paolo, Simone e Giulia, con cui ho trascorso molto tempo e che mi hanno motivato soprattutto negli ultimi mesi.

Infine, un enorme ringraziamento lo devo alla mia famiglia, che mi ha sempre sostenuto e spronato lungo la strada intrapresa.

Sebbene questo elaborato rappresenti di per sé un lavoro conclusivo, preferisco intenderlo come punto di passaggio per il nuovo percorso che mi aspetta.

Marco Cominelli
Brescia, 20 luglio 2016

Indice

Introduzione **1**

1 Tipologie di inquadratura **3**

1.1 Tassonomia delle inquadrature 4

1.2 Stato dell'arte 5

2 Classificazione delle inquadrature cinematografiche **9**

2.1 Schema generale 10

2.2 Estrazione dei descrittori 12

2.3 Apprendimento e classificazione 20

3 Risultati sperimentali **27**

3.1 Preparazione dei dati 28

3.2 Valutazione dei risultati 29

3.2.1 Metodo random 30

3.2.2 Metodo cross-validation 32

3.2.3 Prestazioni dei singoli descrittori 37

3.2.4 Addestramento e test su gruppi di film 40

3.3 Analisi degli errori 42

Conclusione **45**

Bibliografia **47**

Introduzione

La sensazione che film diversi possano essere identificati attraverso proprietà stilistiche comuni ed essere raggruppati secondo criteri come genere, periodo o forme narrative è un'esperienza comune. Individuare e catalogare singoli film come parte di insiemi diversi può essere importante per scopi promozionali, terapeutici o di individuazione delle preferenze dello spettatore.

Esistono anche alcune tecniche facilmente riconoscibili che giocano un ruolo importante nell'impatto emotivo del film sullo spettatore. Gli autori sono consci di queste tecniche e spesso nei loro film le pianificano in modo meticoloso. La scelta del tipo di inquadratura, e in particolare la distanza apparente tra la macchina da presa e il soggetto, è senza dubbio uno dei principali fattori che contribuiscono a determinare le proprietà stilistiche di una scena e di un film.

Recenti studi cinematografici mostrano che, in alcuni casi, l'analisi statistica della distribuzione delle inquadrature (*Shot Scale Distribution*, SSD) rivela schemi ricorrenti nel lavoro di un autore. Per esempio in [1] Kovács ha scoperto una variazione sistematica della SSD nei film di Michelangelo Antonioni, sollevando diversi interrogativi sulle possibili cause estetiche e cognitive di tale regolarità. Sorgono quindi alcune domande che richiedono una risposta: gli schemi delle SSD sono spesso simili nei film dello stesso autore? In quanti casi SSD simili appartengono a film realizzati da autori diversi? Possono le SSD essere considerate una firma dell'autore stesso?

Per rispondere a queste e altre domande, è necessario sviluppare un metodo per la stima automatica delle SSD, dal momento che valutare manualmente la distribuzione delle inquadrature per un intero film è un compito estremamente laborioso. In questo lavoro si studiano le prestazioni del siste-

ma proposto da Benini et al. in [2] proprio per questo scopo. L'approccio utilizzato è quello di analizzare i fotogrammi estratti dal film a intervalli regolari di un secondo. Per ogni fotogramma, le caratteristiche dell'inquadratura vengono sintetizzate da nove descrittori ottenuti combinando diverse tecniche di analisi delle immagini. Lo studio e l'elaborazione di questi descrittori per determinare il tipo di inquadratura sono effettuati attraverso metodi di apprendimento automatico. In particolare, in questo lavoro viene utilizzato un classificatore Random Forest per classificare primi piani (*close-up*, CU), campi medi (*medium shot*, MS) e campi lunghi (*long shot*, LS).

L'obiettivo della tesi è confermare la validità del sistema proposto in [2] confrontando i risultati della stima automatica della SSD con i riferimenti ottenuti da un'analisi manuale. Per fare questo, sono stati condotti dei test su un insieme di dieci film appartenenti alla filmografia di Federico Fellini.

Questo documento si articola nel seguente modo: nel primo Capitolo vengono presentati il problema del riconoscimento delle inquadrature e lo stato dell'arte in questo campo di ricerca; nel secondo Capitolo vengono esposte le tecniche utilizzate per l'analisi e la classificazione automatica delle inquadrature, con approfondimenti sulla rilevanza dei descrittori impiegati e sul funzionamento del classificatore Random Forest; infine, nel terzo Capitolo sono presentati e commentati i risultati ottenuti nella classificazione delle tre categorie principali (LS, MS e CU).

Capitolo 1

Tipologie di inquadratura

Il riconoscimento e la classificazione di inquadrature sono un problema che per essere affrontato richiede l'interconnessione di discipline diverse. In seguito verrano presentati metodi per l'analisi delle proprietà delle immagini e un algoritmo di apprendimento automatico utilizzato per la classificazione delle inquadrature.

In questo Capitolo vengono affrontati i concetti introduttivi per una classificazione corretta delle inquadrature. In particolare, nella Sezione 1.1 vengono discusse le inquadrature cinematografiche più comuni e i criteri utilizzati convenzionalmente per distinguerle, mentre nella Sezione 1.2 sono presentati i possibili approcci a questo problema e lo stato dell'arte nei contesti applicativi.

1.1 Tassonomia delle inquadrature

Nel linguaggio cinematografico, l'inquadratura rappresenta l'unità grammaticale elementare ed è definita come quella porzione di spazio vista attraverso l'obiettivo della macchina da presa. Si può dire che nella narrazione cinematografica tutti gli eventi vengono narrati per mezzo di una o più inquadrature. Storicamente, già nelle prime teorie del linguaggio cinematografico vengono prese in considerazione la pianificazione e la scelta del tipo di inquadratura, in quanto riconosciute come tecniche peculiari per trasmettere informazioni narrative e indurre un coinvolgimento emotivo nello spettatore [3].

Quando ci si avvicina al problema della classificazione tassonomica delle inquadrature e si cerca di definire un approccio sistematico ad essa, ci si scontra necessariamente con una serie di problemi che minano la rigorosità del metodo. Volendo infatti classificare tutte le possibili tipologie di inquadratura, ci si accorge subito di come le possibilità siano pressoché infinite. Inoltre, diversi autori propongono una serie di criteri lassi e soggettivi che spesso dipendono dal loro bagaglio culturale personale.

In questo lavoro ci limitiamo semplicemente a distinguere la distanza apparente tra soggetto e macchina da presa, che è uno dei criteri più comuni utilizzati per classificare i diversi tipi di inquadratura. La distanza è apparente in quanto non dipende in modo univoco dalla distanza reale (fisica) tra soggetto e cinepresa, ma varia anche in funzione di altri fattori, fra cui l'angolazione della ripresa e l'ottica utilizzata.

Sebbene le possibili distanze tra la macchina da presa e il soggetto siano infinite, nei casi pratici la scala delle inquadrature (in inglese *shot scale*) può essere suddivisa in tre categorie principali: campi lunghi (*long shot*, LS), campi medi (*medium shot*, MS) e primi piani (*close-up*, CU). In letteratura è usanza comune distinguere le inquadrature in "piani" e "campi" in base all'importanza del soggetto inquadrato rispetto all'ambiente che lo circonda.

Tenendo presente che in un film esistono anche altri metodi per regolare il coinvolgimento emotivo dello spettatore, si può accettare come regola generale che più l'inquadratura è vicina al soggetto, più la risposta emotiva dello spettatore sarà forte.

(a) (b) (c)

Figura 1.1 - Esempi di a) primi piani, b) campi medi e c) campi lunghi tratti dal film *Otto e mezzo* (1963) di Federico Fellini.

Un primo piano mostra una piccola parte della scena, concentrando l'attenzione sui sentimenti o sulla reazione di una persona, oppure su un dettaglio importante della storia, come in Figura 1.1(a). In questo caso il coinvolgimento dello spettatore è massimo. Definiamo invece campi medi quelle inquadrature in cui gli attori e lo spazio circostante hanno la stessa importanza e quindi occupano approssimativamente lo stesso spazio all'interno della scena, come in Figura 1.1(b). Infine, i campi lunghi mostrano in modo prevalente l'ambiente che circonda il soggetto, come si vede in Figura 1.1(c). La loro funzione è quella di contestualizzare la narrazione in un ambiente oppure di allontanare lo spettatore dalla scena. Nei campi lunghi la partecipazione emotiva dello spettatore è solitamente ridotta.

Come si vede, nella definizione delle varie tipologie di inquadratura ci si scontra necessariamente con l'impiego di termini qualitativi che spesso rendono la distinzione approssimata e poco chiara, un problema di cui si dovrà tenere conto nella valutazione dei risultati proposti in questo lavoro.

1.2 Stato dell'arte

Fino ad ora, solo pochi lavori hanno trattato direttamente il problema della classificazione di inquadrature cinematografiche e la stima della loro distribuzione in un film. Uno di questi, proposto da Cherif et al. in [4], definisce

delle regole basate sulla presenza di una figura umana nella scena per estrarre informazioni sulla SSD su un insieme limitato di 66 fotogrammi, fornendo una descrizione quantitativa dei tipi di inquadratura più comuni. Questo approccio utilizza solamente due descrittori (altezza della testa e distanza dal lato inferiore dello schermo) ed un certo numero di soglie per la classificazione del piano di ripresa. Un nuovo approccio degli stessi autori, presentato in [5], prova a classificare le inquadrature in sette tipologie usando altre due *feature*: il rapporto tra l'altezza del viso e l'altezza dello schermo e il rapporto tra le corrispondenti larghezze. I vettori bidimensionali corrispondenti ai descrittori estratti vengono quindi processati da una macchina a vettori di supporto (SVM) per la classificazione. A causa del tipo di descrittori scelti, in entrambi i casi non è possibile prendere alcuna decisione quando non ci sono attori nell'inquadratura.

Il lavoro di Wang e Cheong in [6] estrapola informazioni sulla distanza di ripresa a partire dal centro dell'attenzione, utilizzandole per distinguere piani e campi in base al rapporto soggetto/sfondo. Sebbene non siano fornite indicazioni sulle prestazioni del sistema, gli autori propongono un approccio sistematico al problema basandosi su descrittori di moto dei fotogrammi. La quantità di dati utilizzati in questo caso è considerevole, anche se il classificatore distingue solamente due tipi di inquadrature (primi piani e campi lunghi), senza considerare le differenze tra un primo piano e un campo medio. Una classificazione basata sul movimento della scena è utilizzata anche in [7] per classificare le inquadrature in otto categorie (aerea, dall'alto, gru, carrello, fissa, panoramica orizzontale/verticale e zoom).

Il lavoro in [8] invece propone un sistema per la classificazione delle inquadrature includendo anche tipologie meno tradizionali come le inquadrature *Over-the-shoulder* e i *Foreground shot*.

Approcci di diverso tipo sono utilizzati nella classificazione semantica delle scene di video sportivi, dove il tipo di inquadratura è, dal punto di vista narrativo, meno importante che nei film. Il sistema proposto in [9] è in grado di distinguere i momenti di "gioco" dalle "pause" valutando la quantità di erba presente nell'inquadratura: i campi lunghi inquadrano una grande area di prato mentre i primi piani hanno aree minori.

Approcci alternativi per determinare il piano di ripresa possono basarsi sulla profondità *assoluta* della scena (ovvero la distanza reale tra la cinepresa e il soggetto). Ci sono diversi esempi di questo tipo in letteratura e tutti si affidano a un limitato numero di informazioni come la visione binoculare, il movimento di parallasse o la variazione del fuoco. Anche in assenza di queste informazioni, una persona sarebbe comunque in grado di stimare in modo approssimativo la profondità assoluta della scena. Nell'approccio proposto da Torralba e Oliva in [10], questa stima viene invece effettuata analizzando le proprietà spettrali della scena inquadrata.

Quando mancano indizi sulla profondità assoluta, la distanza tra l'osservatore e la scena non può essere stimata con grande precisione, a meno che non si riconoscano nell'inquadratura oggetti di dimensioni note [11]. Sfortunatamente, il carico di lavoro richiesto per la segmentazione delle immagini e il riconoscimento di oggetti ha un costo computazionale troppo elevato rispetto all'affidabilità della classificazione.

In ogni caso, la definizione dei principali tipi di inquadratura (LS, MS e CU) non implica necessariamente la valutazione della profondità *assoluta*. Per determinare il piano di ripresa può essere utile stimare solo la distanza *relativa* tra gli elementi della scena. Le diverse tecniche attualmente disponibili per stimare la profondità relativa in un'inquadratura si focalizzano sull'analisi delle ombre [12], dei gradienti delle superfici [13], dei bordi [14], degli schemi geometrici [15] e di altri elementi particolari come le dimensioni relative e l'altezza rispetto alla linea dell'orizzonte [16]. L'interpretazione di ombre, bordi e linee può essere usata per una ricostruzione tridimensionale della scena come in [17], sebbene questi elementi presi singolarmente portino poche informazioni riguardo al tipo di inquadratura.

Questa breve panoramica sulle tecniche e sugli approcci maggiormente utilizzati per affrontare il problema in esame ci permette di definire il contesto in cui va a operare il sistema proposto.

Capitolo 2

Classificazione delle inquadrature cinematografiche

In questo capitolo viene illustrata in dettaglio la struttura del sistema utilizzato per la classificazione delle inquadrature cinematografiche e la stima della loro distribuzione (SSD) in un film. Questo metodo è stato presentato da Benini et al. in [2] e combina l'analisi di più descrittori (*feature*) facilmente estraibili da un film per ottenere una classificazione robusta delle inquadrature e una stima affidabile della sua SSD.

Il sistema può lavorare con diversi algoritmi di apprendimento automatico (*machine learning*), ma questo lavoro si limita a studiarne le prestazioni con il classificatore Random Forest.

La Sezione 2.1 fornisce una panoramica d'insieme sul sistema e ne introduce la struttura; nella Sezione 2.2 viene sviluppato un approfondimento sull'interpretazione dei descrittori utilizzati; nella Sezione 2.3 vengono infine presentati concetti base di apprendimento automatico, soffermandosi in particolare sulle tecniche di apprendimento supervisionato e sull'algoritmo Random Forest.

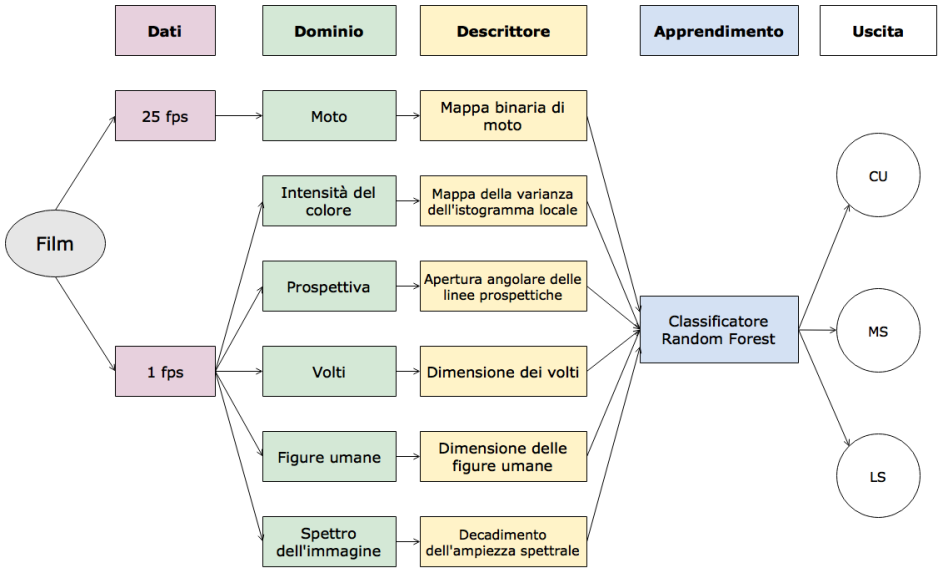


Figura 2.1 - Diagramma di flusso del sistema analizzato per la classificazione delle inquadrature (CU=primo piano, MS=campo medio, LS=campo lungo).

2.1 Schema generale

Il diagramma a blocchi del sistema analizzato è mostrato in Figura 2.1. Come si può vedere, il sistema si compone di più parti che verranno presentate in dettaglio in questo capitolo.

Il sistema classifica il tipo di inquadratura su fotogrammi estratti a intervalli regolari di un secondo. Questa tecnica sembra utilizzare una quantità di dati notevole se confrontata con sistemi simili allo stato dell'arte, ma comporta anche due importanti vantaggi. Per prima cosa, un'analisi "secondo per secondo" permette di classificare le inquadrature anche quando la macchina da presa è in movimento nella scena. Questa è un'eventualità piuttosto comune nel cinema moderno e un'analisi "scena per scena" risulterebbe piuttosto sconveniente e problematica. Il secondo importante vantaggio che consegue dalla tecnica utilizzata consiste nel fatto che, insieme alla SSD del film, il sistema fornisce anche la sequenza delle inquadrature, ciascuna con la propria durata e nell'ordine esatto. Basandosi su questi dati, è possibile generare una rappresentazione temporale del film che può essere confrontata con i dati di altre misurazioni temporali (ad esempio le reazioni emotive o il

livello di attenzione dello spettatore) per qualsiasi tipo di ricerca che indaghi il processo di interazione di un osservatore con l'opera cinematografica.

Per quanto riguarda l'analisi delle inquadrature, il sistema esamina per ogni fotogramma estratto dal film sei proprietà intrinseche della scena rappresentata e le riassume in un vettore di descrittori. Le informazioni ottenute a partire da questi descrittori verranno poi elaborate attraverso tecniche di apprendimento automatico supervisionato.

Il sistema opera direttamente sui fotogrammi estratti dal video, senza la necessità di confrontare diverse immagini della stessa scena per ottenere informazioni spaziali. Questo approccio rimane valido su un insieme eterogeneo di generi cinematografici e non è limitato, come molti altri metodi, all'analisi nel dominio sportivo, dove le informazioni sul colore permettono una classificazione più semplice.

Anziché valutare i piani di ripresa su titoli generici trovati sull'Internet Movie Database [18], si è scelto di seguire il percorso degli autori in [2] e di verificare la validità del sistema su un insieme più impegnativo di film. I film d'autore, a causa delle varietà di sperimentazioni estetiche, della ricchezza delle composizioni sceniche e della presenza di contenuto altamente simbolico [19], possono essere considerati come il materiale più difficile per l'analisi automatica di un film. Al contrario, i comuni prodotti audiovisivi per la televisione e il cinema utilizzano tecniche più convenzionali per la composizione e la scelta delle inquadrature.

In questo lavoro, la verifica sperimentale delle prestazioni del sistema è stata condotta su dieci film appartenenti alla filmografia di Federico Fellini (1920-1993). Fellini è stato uno dei più importanti registi italiani ed è considerato uno dei maggiori protagonisti nella storia del cinema: oltre alla vittoria di numerosi premi in ambito nazionale ed internazionale (fra cui quattro Premi Oscar al miglior film straniero e la Palma d'Oro al Festival di Cannes), sono molti gli autori che si sono ispirati alla sua arte per realizzare nuove opere anche in contesti diversi da quello cinematografico. Per il suo stile visionario e la peculiarità della sua produzione artistica, i suoi film rappresentano un ottimo tipo di materiale per il lavoro che ci siamo proposti. I risultati sono riportati nel Capitolo 3.

2.2 Estrazione dei descrittori

Il metodo proposto per determinare il piano di ripresa utilizza sei tecniche per analizzare il contenuto dei fotogrammi analizzati, contenente informazioni dirette e indirette sulla distanza della macchina da presa dal soggetto della ripresa. I descrittori da cui vengono ricavate queste informazioni sono l'*intensità del colore*, il *movimento*, la *prospettiva*, la presenza di *volti* e *figure umane* e la *rappresentazione spettrale* della scena.

Più precisamente, la prima tecnica investiga la varianza dell'istogramma dell'intensità luminosa calcolata sulle varie regioni di ogni fotogramma. La seconda tecnica lavora su segmenti di dati video e stima la presenza di oggetti in moto in primo piano utilizzando un noto algoritmo di sottrazione dello sfondo. Il terzo metodo investiga la geometria della scena, determinando le linee prospettiche dell'immagine attraverso la trasformata di Hough. Il quarto e il quinto metodo forniscono indicatori sul contenuto della scena, individuando la presenza di volti e di corpi umani, le dimensioni dei quali forniscono una misura indiretta della distanza assoluta tra la macchina da presa e i soggetti ripresi. Infine, ispezionando nel dominio delle frequenze il contenuto spettrale della scena, è possibile distinguere la struttura delle immagini e le loro scale spaziali.

Una volta combinati insieme, i descrittori proposti verrano processati attraverso l'algoritmo di apprendimento automatico supervisionato Random Forest.

Intensità del colore

Quando guardiamo un'immagine, gli elementi distanti non sono nitidi e definiti come quelli in primo piano, a causa della diffusione e diffrazione della luce in un mezzo opaco come l'aria. Le figure più lontane dalla macchina da presa sono più sfocate e la loro intensità uniforme; al contrario, nelle aree dove le figure sono più definite (cioè quelle più vicine all'obiettivo) i livelli di intensità sono più discontinui [20].

Questo descrittore viene calcolato su fotogrammi estratti ad intervalli di un secondo e misura la statistica al secondo ordine degli istogrammi loca-

li dell'immagine. Per prima cosa ogni fotogramma di dimensione $\mathcal{W} \times \mathcal{H}$ viene convertito nella corrispondente immagine $I(x, y)$ indicante l'intensità luminosa. L'istogramma locale dell'intensità viene quindi calcolato su una finestra rettangolare scorrevole w_I di dimensione $\frac{\mathcal{W}}{\mathcal{R}} \times \frac{\mathcal{H}}{\mathcal{R}}$ (dove $\mathcal{R} = 20$, ma è un parametro che può essere variato), centrata attorno al pixel $(\bar{x}, \bar{y})$, che scorre lungo $I(x, y)$.

Indicando con $f(g, w_I)$ il numero di pixel nella finestra w_I il cui livello di grigio è uguale a g, il livello di grigio medio dell'istogramma locale calcolato sulla finestra w_I è calcolato come:

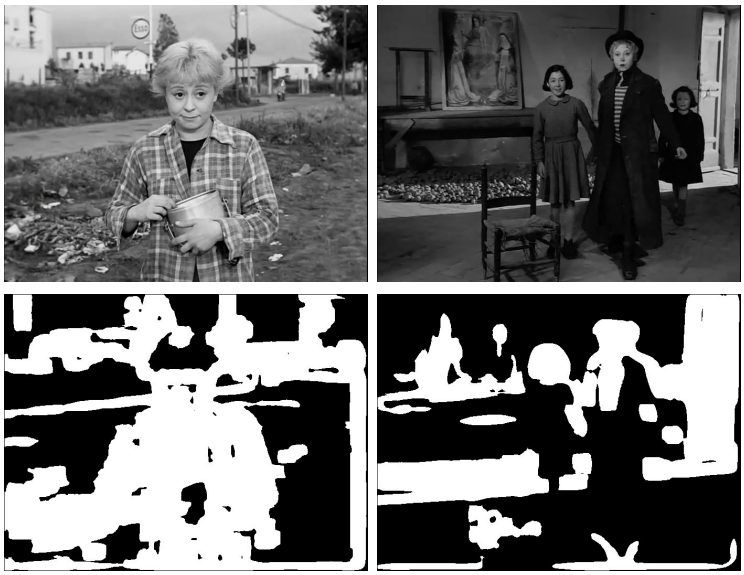
$$\bar{g}_{w_I(\bar{x},\bar{y})} = \frac{\sum_i g_i \cdot f(g_i, w_I)}{\sum_i f(g_i, w_I)} \tag{2.1}$$

e la sua varianza $\sigma^2_{w_I(\bar{x},\bar{y})}$ è calcolata come:

$$\sigma^2_{w_I(\bar{x},\bar{y})} = \frac{\sum_i \left(g_i - \bar{g}_{w_I(\bar{x},\bar{y})}\right)^2 \cdot f(g_i, w_I)}{\sum_i f(g_i, w_I)} \tag{2.2}$$

Troviamo quindi $I_\sigma(x, y)$, un'immagine a livelli di grigio dove il pixel $(\bar{x}, \bar{y})$ rappresenta la varianza $\sigma^2_{w_I(\bar{x},\bar{y})}$ dell'istogramma calcolato sulla finestra scorrevole centrata in $(\bar{x}, \bar{y})$. I valori della varianza vengono quindi normalizzati rispetto al numero totale di fotogrammi. Le parti della scena più distanti dalla macchina da presa saranno le regioni più scure (bassa varianza dell'intensità), mentre le zone più vicine all'osservatore saranno più luminose (varianza maggiore).

Il descrittore scalare è ottenuto attraverso un processo di binarizzazione con una soglia adattabile, come proposto in [21], che classifica ogni pixel del fotogramma come *sfondo* oppure *primo piano*. Dopo aver escluso le componenti più piccole, viene calcolato il rapporto A_σ fra tutti i pixel dello sfondo e i pixel totali (Figura 2.2). Anche se non è possibile stabilire una stretta proporzionalità tra A_σ e il tipo di inquadratura, questo descrittore consente una prima discriminazione grossolana: i campi lunghi avranno un'area di sfondo maggiore, i campi medi minore, i primi piani quasi nulla.



(a) (b)

Figura 2.2 - Esempi del calcolo della varianza dell'istogramma locale nel caso di a) un primo piano e b) un campo medio tratti dal film *La strada* (1954).

Moto

Gli oggetti che si muovono in primo piano sono in genere responsabili di un'alta attività di moto nella scena. Al contrario, un oggetto che si muove in un campo lungo, a causa delle sue dimensioni ridotte, non contribuisce in modo significativo al moto della scena.

Per generare una mappa di moto in grado di descrivere gli oggetti in moto in una scena, si utilizza l'algoritmo proposto da Barnich e Van Droogenbroeck in [22]. Il semplice concetto di sottrazione dello sfondo richiede il confronto dei diversi pixel tra lo sfondo statico e il fotogramma analizzato [23]. Il lavoro in [22] estende questo concetto registrando, per ogni pixel, un insieme di valori assunti in passato dallo stesso pixel o da quelli adiacenti. Per evitare anomalie ed errori, viene utilizzato un modello per un insieme di campioni invece che un modello specifico per i pixel. Di conseguenza, il valore del pixel analizzato è confrontato con quello dei campioni più vicini all'interno della collezione dei campioni e classificato come *primo piano* o *sfondo*.

Mentre gli altri descrittori utilizzati in questo lavoro vengono estratti su

(a) (b)

Figura 2.3 - Mappe di moto a) in un primo piano e b) in un campo lungo. I fotogrammi appartengono al film *La voce della luna* (1990).

singoli fotogrammi, la valutazione del moto è effettuata sui dati video. Il modello dello sfondo viene aggiornato per ogni fotogramma, così da ottenere ogni secondo una mappa binaria del moto nell'inquadratura (bianco per il primo piano e nero per lo sfondo). Dopo una correzione della mappa che escluda le aree più piccole, il descrittore finale è ottenuto misurando il rapporto A_M dell'area bianca rispetto all'area totale dell'inquadratura.

In Figura 2.3 sono presentati esempi di mappe di moto da *La voce della luna*. Pur non distinguendo il moto della macchina da presa da quello degli oggetti ripresi, il descrittore A_M fornisce comunque un'indicazione sulla quantità di oggetti in moto in primo piano. È chiaro che non esiste un rapporto esatto tra la percentuale degli oggetti che si muovono in primo piano e la distanza della macchina da presa, tuttavia questo descrittore può fornire un'indicazione del piano di ripresa: i LS avranno grandi aree di sfondo, mentre i CU saranno perlopiù in primo piano.

(a) (b)

Figura 2.4 - Individuazione delle linee prospettiche a) in un primo piano e b) in un campo lungo nel film *Lo sceicco bianco* (1952).

Prospettiva

Una scena rappresentata in prospettiva presenta sempre uno o più punti di fuga verso cui le linee parallele sembrano convergere. La trasformata di Hough è già stata usata in [24] per individuare i punti di fuga all'interno di un'immagine, ma il compito rimane comunque arduo a causa dell'infinita varietà di scene possibili. Piuttosto che individuare con precisione i punti di fuga della scena, puntiamo invece a stimare la distanza della macchina da presa considerando la pendenza delle linee prospettiche all'interno dell'inquadratura. Dal momento che tutte le linee parallele nella scena convergono verso il punto di fuga, le linee prospettiche in un campo lungo rimangono quasi parallele come in Figura 2.4(b). Al contrario, l'inclinazione delle linee prospettiche è evidentemente diversa nei primi piani, come si vede in Figura 2.4(a).

Per prima cosa, le possibili linee prospettiche vengono individuate nella scena da un *edge detector* con l'operatore di Canny. Secondo la trasformata di Hough probabilistica, ogni punto nell'immagine viene mappato in un luogo di punti nel piano di Hough che corrispondono a tutte le linee rette passanti per quel punto. Sommando tutti i contributi, le linee che appaiono nell'immagine originale sono i massimi locali nel piano di Hough. Misurando gli angoli di inclinazione delle linee prospettiche rispetto all'asse verticale $\bar{u}_y$, si è in grado di ottenere uno stimatore della distanza della cinepresa. Indicando con θ_i

l'angolo delle n linee prospettiche il cui angolo con la verticale è compreso nell'intervallo $0 < \theta_i < \pi/2$, e con ϕ_k l'angolo delle m linee prospettiche il cui angolo con la verticale è nell'intervallo $\pi/2 < \phi_k < \pi$, le inclinazioni medie $\bar{\theta}$ e $\bar{\phi}$ sono:

$$\bar{\theta} = \frac{1}{n}\sum_{i=1}^{n}\theta_i \quad , \quad \bar{\phi} = \frac{1}{m}\sum_{k=1}^{m}\phi_k \quad . \tag{2.3}$$

Le linee orizzontali e verticali vengono ignorate, dal momento che non portano alcuna informazione sulla struttura prospettica della scena. L'*apertura angolare* α delle linee prospettiche è data dalla differenza tra le due inclinazioni medie, ovvero:

$$\alpha = \frac{\bar{\phi} - \bar{\theta}}{2} \tag{2.4}$$

Questo descrittore ci fornisce un'indicazione di massima sulla distanza tra la macchina da presa e la scena inquadrata.

Volti

Ad eccezione di pochi film, la presenza di figure umane è un elemento centrale nella filmografia moderna e la probabilità che un volto sia mostrato in una scena ha un'importanza rilevante nel determinarne il contenuto. A supporto di questa considerazione, il lavoro di Tsingalis et al. in [5] ha già provato a stimare il piano di ripresa basandosi sulle dimensioni dei volti inquadrati. Per garantire un buon tasso di riconoscimento anche su viste non frontali, il nostro sistema utilizza il metodo proposto in [25] da Zhu e Ramanan, il quale risulta essere molto efficace anche nel catturare deformazioni elastiche dei volti.

Le prestazioni di questo classificatore sono molto buone considerando che il modello è addestrato con poche centinaia di volti mentre alcuni sistemi commerciali utilizzano anche miliardi di campioni. Sebbene il metodo utilizzato restituisca anche una stima della posa e dei riferimenti spaziali, il metodo utilizzato ricava solamente il rapporto fra l'area del volto più grande riconosciuto e la dimensione totale del fotogramma, ottenendo il descrittore

(a) (b)

Figura 2.5 - Esempio di volti individuati a) in un primo piano e b) in un campo medio nel film *E la nave va* (1983).

A_F. È evidente che sia la presenza di un volto sia la dimensione dello stesso forniscono una descrizione indiretta del tipo di inquadratura: le inquadrature con volti più grandi saranno probabilmente dei primi piani, mentre si avranno volti piccoli o addirittura assenti nei campi lunghi. In Figura 2.5 sono forniti esempi in tal senso tratti dal film *E la nave va*.

Figure umane

Partendo dalle considerazioni già svolte per quanto riguarda il riconoscimento dei volti, anche la valutazione delle dimensioni di un'intera figura umana può fornire informazioni importanti sul tipo di inquadratura. Questo è un tema complicato, in quanto l'algoritmo utilizzato dovrebbe prima di tutto essere in grado di riconoscere sia corpi interi che parzialmente occlusi; inoltre dovrebbe poter stimare le dimensioni effettive dei corpi non completamente visibili nell'inquadratura (situazione tipica nei primi piani e nei campi medi); infine, l'ultima abilità richiesta è quella di identificare corpi ripresi da dietro, situazione piuttosto comune nel cinema moderno.

Come soluzione a questo problema, il sistema utilizza l'algoritmo di *object detection* proposto da Felzenszwalb et al. in [26], addestrato per l'identificazione di pedoni.

Una volta individuati tutte le figure umane in un'inquadratura, il descrittore A_B è espresso come il rapporto tra l'area completa (stimata) del corpo più grande e l'area totale del fotogramma. Esempi dei risultati ottenuti sono visibili in Figura 2.6.

(a) (b)

Figura 2.6 - Esempio di individuazione di figure umane a) in un primo piano e b) in un campo lungo nel film *Otto e mezzo* (1963).

Spettro dell'immagine

Come suggerito da Torralba e Oliva in [10], il modulo della trasformata discreta di Fourier (DFT) di un'immagine $I(x, y)$ di dimensione $\mathcal{W} \times \mathcal{H}$:

$$\left|\mathcal{I}\left(\bar{f}\right)\right| = \left|\sum_{x=0}^{\mathcal{W}-1} \sum_{y=0}^{\mathcal{H}-1} I(x, y) e^{-j2\pi\left(\frac{xf_x}{\mathcal{W}} + \frac{yf_y}{\mathcal{H}}\right)}\right| \tag{2.5}$$

contiene informazioni riguardanti sia la natura che la scala dell'immagine.

In particolare, le componenti lungo le direzioni principali (orizzontale, verticale, diagonale) forniscono informazioni rilevanti riguardo al contenuto della scena. Inquadrature che rappresentano scene naturali sono caratterizzate da uno spettro tendenzialmente uniforme. Al contrario, quando vengono rappresentate scene di edifici o con motivi altamente geometrici si ha una prevalenza spettrale delle componenti verticali e orizzontali. In genere si può quindi dire che i primi piani possiedono uno spettro più uniforme.

Gli stessi autori hanno mostrato che anche l'analisi del decadimento dell'ampiezza spettrale lungo le direzioni dominanti è utile per lo studio dell'inquadratura. In scene naturali in cui le asperità dell'immagine diminuiscono con la distanza l'energia si concentra soprattutto alle basse frequenze, mentre le scene geometriche presentano motivi più elaborati. Per investigare la relazione tra le strutture dell'immagine e la distanza D tra la cinepresa e la scena, l'ampiezza spettrale della DFT dell'Equazione 2.5 è modellizzata dagli autori come:

$$\left|\mathcal{I}\left(\bar{f}\right)\right| \sim A(D,\theta)/\|\overline{f}\|^{\Gamma(D,\theta)} \tag{2.6}$$

dove θ è la fase del vettore delle frequenze $\overline{f}$, $A(D,\theta)$ un fattore di grandezza e $\Gamma\left(D,\theta\right)$ la pendenza con cui decade l'ampiezza spettrale.

Il descrittore finale $\overline{\Phi}$ si compone quindi di quattro elementi:

$$\overline{\Phi} = \left[\left|\mathcal{I}(\bar{f})\right|, \gamma^v, \gamma^h, \gamma^d\right] \tag{2.7}$$

dove il primo elemento è il modulo della DFT e gli altri sono indicatori della pendenza dell'ampiezza spettrale lungo le direzioni principali (pendenza verticale γ^v, orizzontale γ^h e diagonale γ^d).

2.3 Apprendimento e classificazione

Al termine del processo di estrazione delle *feature* appena descritto, si ottiene una matrice di valori che rappresenta le proprietà di ogni fotogramma analizzato. A questo punto, il problema si sposta dall'analisi delle immagini all'interpretazione dei dati a disposizione per classificare correttamente l'inquadratura.

Situazioni di questo tipo, in cui una macchina cerca di simulare le funzioni di ragionamento umane, vengono affrontate nel campo dell'intelligenza artificiale e rappresentano un tipico problema di apprendimento automatico (in inglese *machine learning*). L'obiettivo è in generale quello di migliorare le capacità di apprendimento della macchina sulla base di un'esperienza pregressa, ma è necessario dare alcune definizioni prima di procedere con la descrizione dell'algoritmo utilizzato per affrontare il problema.

Le tecniche di *machine learning* si dividono in *classificazione* e *regressione*. I metodi di classificazione cercano di predire l'appartenenza di un oggetto a una certa classe dato un insieme discreto di possibili classi. I primi studi sul problema della classificazione sono stati condotti da Fisher [27] nel 1936 e affrontavano problemi di classificazione binaria (selezione fra due sole classi). Lavori successivi hanno esteso l'analisi a problemi di classificazione multi-classe, utilizzando anche funzioni non lineari [28]. Le tecniche di re-

gressione, al contrario, producono in uscita un risultato che può assumere valori continui (ad esempio un numero reale) e non sono utilizzate in questo lavoro.

In base al tipo di esperienza si è soliti distinguere anche diversi tipi di apprendimento. Nell'*apprendimento supervisionato*, il classificatore viene prima di tutto addestrato con una serie di osservazioni considerate "esatte" a priori in cui compaiono sia i dati di ingresso che l'uscita desiderata (fase di addestramento). Un algoritmo di apprendimento supervisionato può così generare una funzione di inferenza che associa a ogni vettore di dati in ingresso l'uscita più probabile. Nell'*apprendimento non supervisionato* invece vengono forniti solo i dati in ingresso e il compito del sistema è quello di individuare e/o raggruppare certe ricorrenze o similitudini.

Le diverse osservazioni, così come gli oggetti da classificare, vengono spesso rappresentate attraverso delle proprietà quantificabili che vengono riassunte in valori detti descrittori (*feature*). Inoltre, questi descrittori utilizzati per rappresentare l'oggetto possono essere binari o appartenere a un insieme discreto di valori, oppure essere indicati con dei valori reali.

Sulla base di quanto appena visto, possiamo dire che l'approccio utilizzato nel sistema in esame prevede una classificazione multi-classe attraverso un algoritmo di apprendimento supervisionato che fa uso di un vettore di nove elementi di *input* (descrittori) per stimare la classe di uscita (CU, MS, LS).

Classificatori ad alberi

Come detto prima, l'obiettivo del *machine learning* è quello di sintetizzare le proprietà di ragionamento umane all'interno di un algoritmo che permetta ad un calcolatore di prendere decisioni in modo automatico. Una tecnica comune per affrontare questo tipo di problemi è quella di utilizzare degli alberi decisionali.

In generale, un albero k-dimensionale è una struttura utilizzata per organizzare i punti in uno spazio con k dimensioni. Ogni nodo dell'albero è rappresentato da un punto k-dimensionale: i nodi foglia rappresentano le

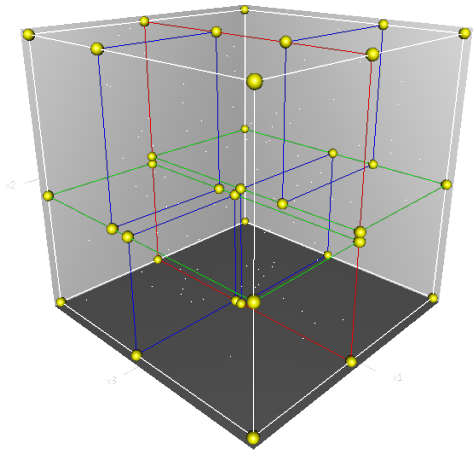


Figura 2.7 - Esempio di albero decisionale a tre dimensioni. Il primo nodo decisionale (piano verticale rosso) divide lo spazio base in due celle, ognuna delle quali è divisa (piani verdi orizzontali) in altre due sottocelle. Infine, queste quattro celle sono ancora divise (piani blu verticali) a due a due. I rimanenti otto nodi ai vertici del cubo rappresentano i nodi foglia, ovvero le possibili classi di appartenenza. L'immagine è tratta da [29].

diverse classi di appartenenza, mentre i nodi decisionali sono generatori di iperpiani che dividono in due lo spazio (Figura 2.7).

Gli alberi decisionali sono strutture che possono essere facilmente percorse fino ai nodi foglia iterando più test sui diversi attributi (Figura 2.8), in modo molto simile a quanto accade con i diagrammi di flusso. A causa della loro semplicità di implementazione e velocità di esecuzione da parte di un calcolatore, rappresentano una delle tecniche dominanti nei problemi di apprendimento automatico. Un albero decisionale può restituire dei risultati anche con un numero ridotto di dati in ingresso ed è versatile per l'analisi combinata con altre tecniche decisionali; tuttavia la presenza di dati incerti e la discesa in profondità lungo l'albero possono causare confusione e uno sbilanciamento verso i dati con un maggior numero di livelli.

Algoritmo Random Forest

Random Forest è un algoritmo sviluppato da Leo Breiman e Adele Cutler. Una presentazione esaustiva di Random Forest è data dagli autori in [31];

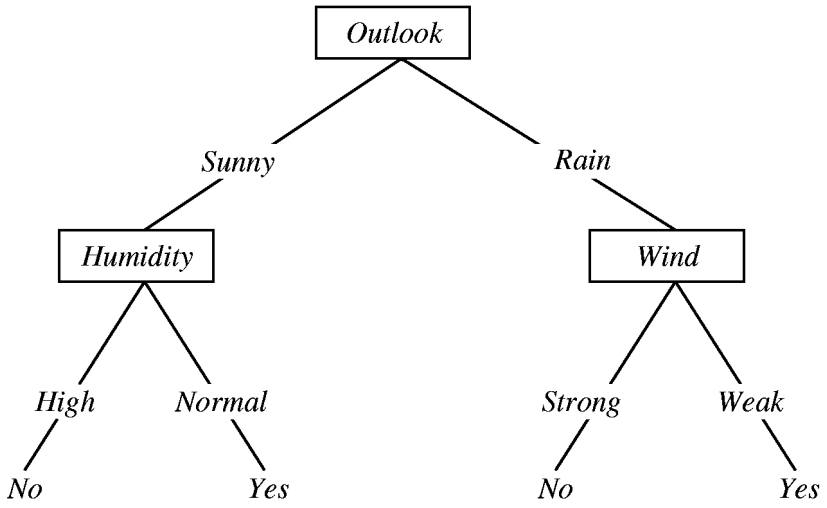


Figura 2.8 - Semplice esempio di albero decisionale che con due sole domande decide se è possibile giocare una partita di tennis [30].

in questo paragrafo si vogliono solo introdurre i concetti fondamentali e commentare le proprietà principali dell'algoritmo.

Come visto prima, gli alberi decisionali sono una tecnica molto utilizzata nel campo del *machine learning* perché producono modelli robusti e ispezionabili, spesso tuttavia risultano poco accurati. In generale, alberi troppo "grandi" rischiano di apprendere schemi non regolari che si adattano in modo eccessivo all'insieme di *training* e falliscono poi in fase di test (*overfitting*). L'algoritmo Random Forest riduce questo problema mediando l'uscita di più alberi decisionali allenati su diverse parti dello stesso *training set* con l'obiettivo di ridurre la varianza.

L'algoritmo di *training* utilizza la tecnica di *bagging* introdotta dallo stesso Breiman per la formazione degli alberi. All'interno del *training set* si seleziona un ulteriore insieme di campioni casuali su cui è fatto crescere un albero decisionale. I campioni rimanenti formano i cosiddetti dati OOB (*out-of-bag*) sui quali viene testato l'albero appena creato. Ripetendo il procedimento più volte con diverse estrazioni dal *training set*, si ottiene un insieme di alberi decisionali (foresta) i cui risultati di classificazione sono pesati in base alle prestazioni nella stima dei dati OOB. La classificazione dei campioni di test verrà poi svolta tenendo conto dei pesi di ogni albero (Figura 2.9).

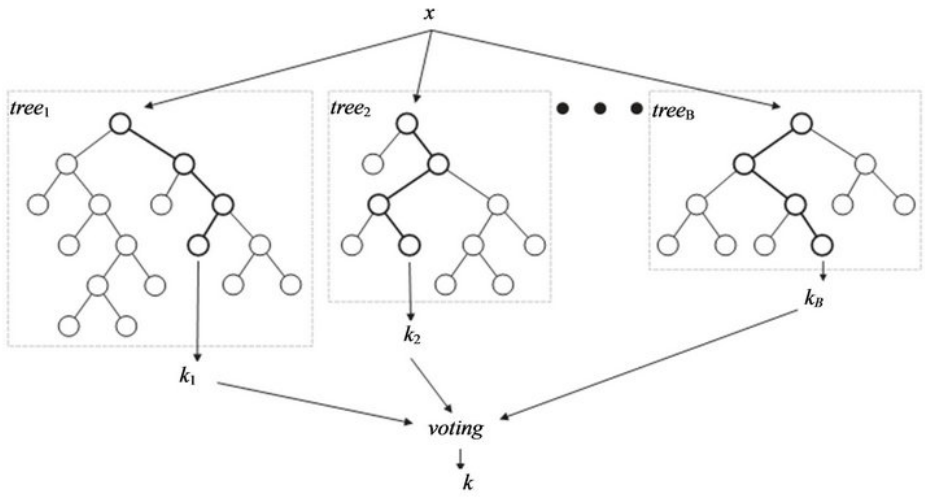


Figura 2.9 - Esempio di classificazione di un elemento x con Random Forest. Ogni albero decisionale può classificare in modo diverso l'elemento in esame. Alla fine la decisione viene effettuata considerando le classificazioni pesate di ogni albero. L'immagine è tratta da [32].

Questa procedura presenta delle performance migliori rispetto a un singolo albero decisionale, perché l'effetto del rumore viene attenuato dalla media di più alberi scorrelati fra loro. Il numero di alberi è arbitrario, ma in genere varia in base alla dimensione dell'insieme di addestramento e può arrivare a diverse migliaia. Il numero ottimo può essere individuato con la tecnica di *cross validation* oppure osservando l'errore nella stima dei dati OOB.

Random Forest aggiunge però un ulteriore livello di casualità alla tecnica di *bagging*. Oltre a costruire un albero diverso per ogni campione casuale estratto dal *training set*, viene cambiato anche il metodo di generazione degli alberi. Se nei classificatori ad alberi ogni nodo viene diviso usando la migliore divisione tra tutte le variabili, in Random Forest ogni nodo viene diviso sulla base del miglior descrittore scelto fra un sottoinsieme casuale dei descrittori utilizzati. In questo modo gli alberi della foresta sono "più diversi" fra loro e la predizione più accurata. Questa strategia rende inoltre la classificazione robusta all'*overfitting*.

Gli sviluppatori dell'algoritmo evidenziano fra le varie peculiarità di Random Forest anche una serie di elementi che lo rendono vantaggioso rispetto a

molti altri algoritmi di classificazione. Fra questi sottolineiamo che Random Forest ci permette di gestire un grande numero di variabili in ingresso ed essere allo stesso tempo molto veloce durante la fase di *learning*.

Capitolo 3

Risultati sperimentali

Nei capitoli precedenti è stato introdotto il problema della classificazione delle inquadrature cinematografiche ed è stato presentato un sistema in grado di svolgere questo compito combinando l'analisi di più descrittori.

Per verificare le prestazioni e l'effettiva validità del sistema proposto, si è scelto di metterlo alla prova con un *dataset* composto da una parte della filmografia del regista italiano Federico Fellini. L'addestramento e il test sono stati effettuati su un campione di alcune decine di migliaia di fotogrammi, confrontando i risultati del test con dei valori di riferimento derivanti da annotazioni manuali.

La Sezione 3.1 descrive la tipologia di dati selezionati e il procedimento utilizzato per ricavarli; nella Sezione 3.2 vengono presentati e commentati i risultati del sistema sul *data set* scelto; infine, nella Sezione 3.3 sono analizzati gli errori di classificazione e sono discusse le principali cause.

Titolo	**Anno**	**Fotogr.**	**CU %**	**MS %**	**LS %**
Lo sceicco bianco	1952	4278	64,0	26,7	9,3
I vitelloni	1953	4483	60,5	25,5	14,0
La strada	1954	4261	54,6	29,5	15,9
Otto e mezzo	1963	6915	69,4	20,7	9,9
Fellini Satyricon	1969	6171	55,8	28,6	15,6
Roma	1972	4420	30,3	48,0	21,7
Amarcord	1973	4211	39,2	28,8	32,0
Il Casanova di Fellini	1976	6689	58,8	26,2	15,0
E la nave va	1983	4453	38,3	29,7	32,0
La voce della luna	1990	4230	47,4	23,1	29,5

Tabella 3.1 - Elenco dei film utilizzati per studiare le prestazioni del sistema.

3.1 Preparazione dei dati

Il *dataset* scelto comprende 10 lungometraggi della filmografia di Federico Fellini, girati tra il 1952 e il 1990. I "valori veri" (*ground truth*) utilizzati per addestrare il sistema e verificarne le prestazioni sono stati annotati manualmente e messi a disposizione dal prof. A.B. Kovács dell'Università ELTE di Budapest.

L'elenco dei film utilizzati in questo lavoro è riportato in Tabella 3.1; per ogni film sono indicati anche l'anno di produzione, il numero di fotogrammi analizzati e la distribuzione reale delle inquadrature (SSD). Utilizzando un insieme di film realizzati in un arco temporale di circa quarant'anni, oltre alle differenze stilistiche tra i singoli film abbiamo a che fare anche con tipologie di materiale chiaramente differenti. In particolare, nel *dataset* rientrano quattro film in bianco e nero, da *Lo sceicco bianco* (1952) a *Otto e mezzo* (1963), e sei a colori, da *Fellini Satyricon* (1969) in poi.

Poiché le annotazioni manuali distinguono nove diversi tipi di inquadrature, è stato necessario adattarle alle tre classi di interesse trascurando alcuni tipi di inquadrature (ad esempio i titoli di testa e di coda o altre tipologie particolari come i *Foreground Shot*, dove tutta la scena è a fuoco). Come conseguenza di questo fatto, il numero di fotogrammi e la SSD dei singoli film potrebbero variare rispetto ai dati indicati dal prof. Kovács.

Per ogni film sono stati estratti i nove descrittori presentati nel Capitolo 2.2 (varianza dell'istogramma dell'intensità luminosa A_σ, apertura angolare delle linee prospettiche α, mappa di movimento A_M, dimensione del volto più grande A_F, dimensione della figura umana più grande A_B, vettore delle proprietà spettrali $\overline{\Phi}$) a intervalli regolari di un secondo, per un totale di oltre 50.000 fotogrammi analizzati.

3.2 Valutazione dei risultati

Si riportano e si commentano ora i risultati ricavati a partire dall'insieme di film in Tabella 3.1. Per la verifica sperimentale delle prestazioni, il sistema è stato sottoposto a diverse condizioni di test.

Per prima cosa, sono stati condotti dei test scegliendo come *training set* e *testing set* insiemi di campioni appartenenti al medesimo film. La selezione dell'insieme di *training* è stata effettuata sia casualmente che con la tecnica di *cross-validation*. In entrambi i casi è stata svolta anche una verifica delle prestazioni dei singoli descrittori, sinteticamente riportata in seguito. Infine si è scelto un insieme di quattro film come *training set* che è stato utilizzato per stimare la SSD degli altri sei film. In tutti i casi le prestazioni sono state valutate con indicatori classici nel campo dell'apprendimento automatico: matrici di confusione, misura F_1, *precision*, *recall* e accuratezza.

Nelle matrici di confusione, l'elemento sulla riga i e sulla colonna j indica la percentuale di "valori veri" di tipo i che sono stati classificati di tipo j dal sistema. Un buon classificatore avrà sulla diagonale della matrice di confusione tutti i valori tendenti a 1.

Con riferimento al grafico in Figura 3.1, *precision* e *recall* nella classificazione di ogni classe sono definite nel seguente modo:

$$\text{Precision} = \frac{veri\ positivi}{veri\ positivi + falsi\ positivi}$$

$$\text{Recall} = \frac{veri\ positivi}{veri\ positivi + falsi\ negativi}$$

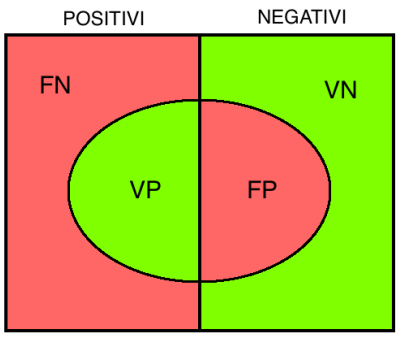


Figura 3.1 - Dato l'insieme di tutti i campioni, il sistema classifica i campioni all'interno dell'ovale come appartenenti alla classe in esame. Possono verificarsi quattro casi: 1) un campione è correttamente classificato come appartenente alla classe (VP - vero positivo), 2) un campione è correttamente classificato come non appartenente alla classe (VN - vero negativo), 3) un campione è erroneamente classificato come appartenente alla classe (FP - falso positivo), 4) un campione è erroneamente classificato come non appartenente alla classe (FN - falso negativo).

La *precision* misura quindi il numero di veri positivi rispetto a tutti i positivi trovati mentre la *recall* misura il numero di veri positivi individuati rispetto a tutti i veri positivi. La *misura* F_1 è ottenuta per ogni classe come media armonica tra *precision* e *recall*, mentre l'*accuratezza* è data dalla percentuale di campioni classificati in modo corretto.

3.2.1 Metodo random

L'approccio utilizzato per ogni film in esame è il seguente: il *training set* si compone scegliendo casualmente il 75% dei campioni appartenenti alla classe con numerosità minore e poi lo stesso numero di campioni per le altre due classi; il test è svolto sui campioni rimanenti dello stesso film.

L'utilizzo di un *training set* bilanciato (ovvero con lo stesso numero di campioni per ogni classe) serve a limitare i problemi legati all'*overfitting*, ma così facendo il sistema presenta delle prestazioni leggermente inferiori rispetto a quelle indicate in [2]. Qui gli autori utilizzano un *training set* non bilanciato basandosi sulla probabilità a priori che un campione appartenga ad una

determinata classe e mediando i risultati su venti ripetizioni dell'esperimento.

I risultati di questo metodo sono riportati a titolo d'esempio nelle Tabelle 3.2-3.6 per cinque tra i film analizzati: *Lo sceicco bianco* (1952), *La strada* (1954), *Otto e mezzo* (1963), *Roma* (1972), *E la nave va* (1983). In particolare, vengono riportate le matrici di confusione, misura F_1, precision, recall e accuratezza per ognuno di questi film.

In Figura 3.2 sono mostrate le prestazioni totali del sistema calcolate su tutti e dieci i film del *dataset* col metodo *random*, indicando media e deviazione standard di precision, recall, misura F_1 e accuratezza.

Inquadratura	CU	MS	LS	
CU	**0,8474**	0,1268	0,0258	
MS	0,2073	**0,7369**	0,0558	
LS	0,1770	0,1610	**0,6620**	
	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	88,35	92,28	84,74	
MS	67,13	61,70	73,69	**81,57 %**
LS	54,23	46,08	66,20	

Tabella 3.2 - Matrice di confusione e prestazioni su *Lo Sceicco Bianco* (1952), metodo *random*.

Inquadratura	CU	MS	LS	
CU	**0,8412**	0,1148	0,0440	
MS	0,0938	**0,8194**	0,0868	
LS	0,0515	0,0674	**0,8812**	
	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	89,26	95,07	84,12	
MS	77,59	73,70	81,94	**83,77 %**
LS	64,52	50,95	88,12	

Tabella 3.3 - Matrice di confusione e prestazioni su *La strada* (1954), metodo *random*.

Inquadratura	CU	MS	LS	
CU	**0,7922**	0,1269	0,0808	
MS	0,1551	**0,7064**	0,1385	
LS	0,0846	0,1116	**0,8038**	
	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	86,64	95,59	79,22	
MS	60,86	53,50	70,64	**78,10 %**
LS	35,48	22,79	80,38	

Tabella 3.4 - Matrice di confusione e prestazioni su *Otto e mezzo* (1954), metodo *random*.

Inquadratura	CU	MS	LS	
CU	**0,7581**	0,1746	0,0673	
MS	0,1938	**0,6464**	0,1598	
LS	0,0618	0,1224	**0,8158**	
	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	68,24	62,07	75,81	
MS	74,11	86,87	64,64	**69,49 %**
LS	55,84	42,46	81,58	

Tabella 3.5 - Matrice di confusione e prestazioni su *Roma* (1972), metodo *random*.

Inquadratura	CU	MS	LS	
CU	**0,8056**	0,1255	0,0689	
MS	0,1038	**0,7573**	0,1390	
LS	0,0528	0,1015	**0,8457**	
	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	85,43	90,95	80,56	
MS	70,11	65,31	75,73	**80,65 %**
LS	81,90	79,41	84,57	

Tabella 3.6 - Matrice di confusione e prestazioni su *E la nave va* (1983), metodo *random*.

3.2.2 Metodo cross-validation

La selezione casuale del *training set* e del *testing set* può portare a problemi di *overfitting* e di prossimità spaziale tra i campioni di addestramento e di test

che falsificano il risultato della verifica svelando prestazioni più ottimistiche rispetto alla realtà.

Per evitare questi problemi, i test sono stati ripetuti utilizzando il metodo *10-fold cross-validation*. Questo metodo prevede la suddivisione dei campioni in dieci parti, utilizzando poi a rotazione nove parti per il *training* e una per il *testing* e mediando i risultati.

Vengono a titolo d'esempio riportati i risultati ottenuti con questo metodo per gli stessi film di prima: *Lo sceicco bianco* (1952), *La strada* (1954), *Otto e mezzo* (1963), *Roma* (1972), *E la nave va* (1983). In particolare, vengono riportate le matrici di confusione, misura F_1, precision, recall e accuratezza per ognuno di questi film.

Inquadratura	CU	MS	LS	
CU	**0,8044**	0,1542	0,0414	
MS	0,2601	**0,6567**	0,0832	
LS	0,2588	0,2789	**0,4623**	
	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	82,46	84,59	80,44	
MS	61,88	58,50	65,67	**73,30 %**
LS	46,58	46,94	46,23	

Tabella 3.7 - Matrice di confusione e prestazioni su *Lo Sceicco Bianco* (1952), metodo *10-fold cross-validation*.

Inquadratura	CU	MS	LS	
CU	**0,7941**	0,1504	0,0556	
MS	0,1480	**0,7430**	0,1090	
LS	0,1369	0,1205	**0,7426**	
	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	82,98	86,89	79,41	
MS	71,27	68,48	74,30	**77,08 %**
LS	69,45	65,23	74,26	

Tabella 3.8 - Matrice di confusione e prestazioni su *La strada* (1954), metodo *10-fold cross-validation*.

Inquadratura	CU	MS	LS	
CU	**0,7604**	0,1404	0,0992	
MS	0,2132	**0,6087**	0,1781	
LS	0,1773	0,2304	**0,5923**	
	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	82,25	89,57	76,04	
MS	55,61	51,18	60,87	**71,25 %**
LS	44,38	35,49	59,23	

Tabella 3.9 - Matrice di confusione e prestazioni su *Otto e mezzo* (1954), metodo *10-fold cross-validation.*

Inquadratura	CU	MS	LS	
CU	**0,6590**	0,2634	0,0776	
MS	0,2791	**0,5277**	0,1931	
LS	0,0998	0,2910	**0,6092**	
	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	60,65	56,17	65,90	
MS	57,87	64,06	52,77	**58,51 %**
LS	56,67	52,97	60,92	

Tabella 3.10 - Matrice di confusione e prestazioni su *Roma* (1972), metodo *10-fold cross-validation.*

Inquadratura	CU	MS	LS	
CU	**0,7015**	0,1848	0,1138	
MS	0,2094	**0,5707**	0,2200	
LS	0,0992	0,1575	**0,7433**	
	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	72,07	74,10	70,15	
MS	57,70	58,35	57,07	**67,60 %**
LS	71,32	68,55	74,33	

Tabella 3.11 - Matrice di confusione e prestazioni su *E la nave va* (1983), metodo *10-fold cross-validation.*

Sebbene inferiori a quelle del metodo *random*, le prestazioni restano co-

munque buone. Media e deviazione standard dei risultati ottenuti con il metodo *10-fold cross-validation* sono mostrate in Figura 3.3.

Non sorprende il fatto che le prestazioni per le singole classi siano sbilanciate in favore della classe con la popolazione maggiore. Ad esempio, risulta particolarmente evidente ne *Lo sceicco bianco* (1952) un netto divario nei risultati ottenuti sui CU e sui LS. Ciò dipende dal fatto che il *training set* utilizzato non è bilanciato e solo il 9,3% delle inquadrature sono campi lunghi. Nei film dove la distribuzione delle inquadrature è più uniforme, anche i risultati sono più equilibrati fra le diverse classi.

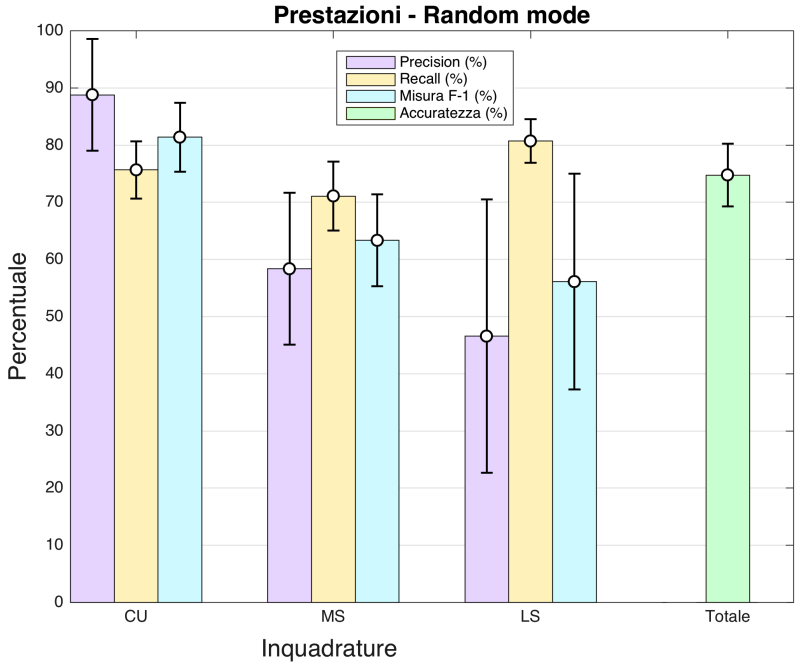


Figura 3.2 - Prestazioni del sistema sui singoli film (media e deviazione standard), metodo *random*.

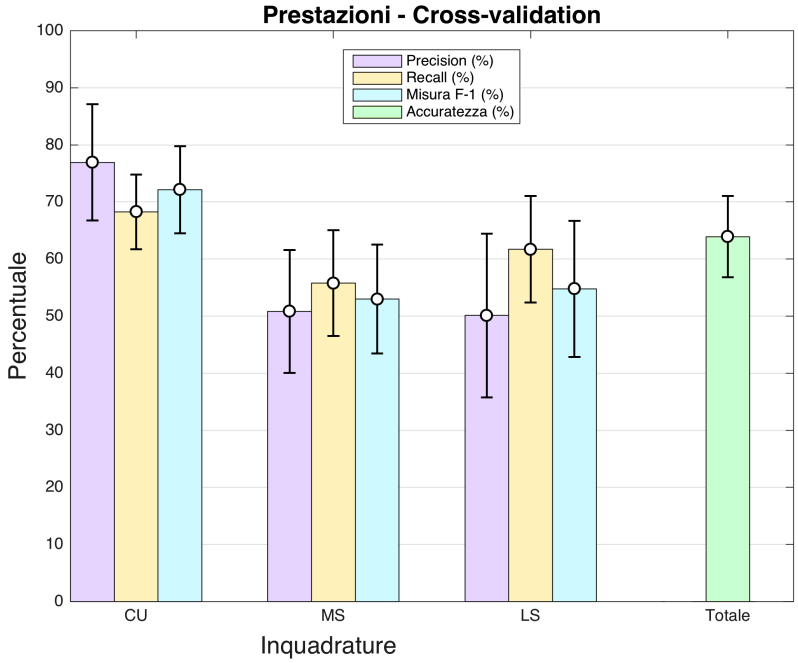


Figura 3.3 - Prestazioni del sistema sui singoli film (media e deviazione standard), metodo *10-fold cross-validation*.

3.2.3 Prestazioni dei singoli descrittori

Il lavoro di analisi è stato esteso anche alle prestazioni del sistema addestrato con i singoli descrittori. Sebbene risulti chiaro che la combinazione di più *feature* migliori le prestazioni complessive del sistema, è comunque interessante cercare di capire la capacità di ogni descrittore di distinguere le diverse inquadrature.

Nelle Tabelle 3.12 e 3.13 sono riportati i risultati sul film *I vitelloni* (1953) utilizzando prima tutti i descrittori e poi solo uno per volta. I risultati dell'analisi dei singoli descrittori sugli altri film sono stati omessi in quanto presentano un comportamento molto simile.

Inquadratura	CU	MS	LS	
CU	**0,6973**	0,2207	0,0819	
MS	0,2818	**0,5250**	0,1932	
LS	0,1415	0,2090	**0,6495**	
	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	75,45	82,20	69,73	
MS	48,52	45,10	52,50	**64,68 %**
LS	55,04	47,75	64,95	

Tabella 3.12 - Prestazioni del sistema sul film *I vitelloni* (1953) utilizzando tutti i descrittori, metodo *10-fold cross-validation.*

Da una rapida valutazione dei risultati emerge una preponderanza dei descrittori legati alla presenza umana: i descrittori dei volti (A_F) e dei corpi (A_B) presentano le migliori prestazioni di classificazione rispetto ai criteri di valutazione utilizzati. Un volto o un corpo identificati all'interno di un fotogramma sono un buon indizio sulla distanza assoluta tra il soggetto e la macchina da presa, in quanto hanno una dimensione abbastanza nota. Tuttavia, sebbene la probabilità di trovare un volto in una scena è altamente legato alla tipologia di film, non è sempre garantito che la presenza dei volti venga rilevata nelle inquadrature. Per esempio nei campi lunghi le persone sono talvolta inquadrate da grandi distanze e in questo caso i volti, a causa delle loro dimensioni ridotte, sono difficilmente individuabili.

		A_σ (Intensità del colore)		
Inquadratura	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	45,20	62,28	35,47	
MS	29,51	26,29	33,63	**34,83 %**
LS	20,36	14,49	34,24	
		A_M (Movimento)		
Inquadratura	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	45,92	64,23	35,73	
MS	31,33	27,34	36,70	**36,09 %**
LS	22,08	15,83	36,50	
		α (Prospettiva)		
Inquadratura	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	47,67	65,63	37,43	
MS	32,00	28,64	36,26	**37,34 %**
LS	22,99	16,32	38,91	
		A_F (Volti)		
Inquadratura	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	59,33	68,47	52,34	
MS	32,93	28,45	39,07	**43,76 %**
LS	12,77	11,14	14,95	
		A_B (Figure umane)		
Inquadratura	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	71,80	82,54	63,53	
MS	45,96	42,19	50,48	**60,56 %**
LS	50,00	40,22	66,08	
		$\overline{\Phi}$ (Spettro dell'immagine)		
Inquadratura	F_1 (%)	Prec. (%)	Rec. (%)	**Accuratezza**
CU	51,20	65,00	42,23	
MS	28,78	25,40	33,19	**37,70 %**
LS	17,68	13,34	26,21	

Tabella 3.13 - Prestazioni del sistema sul film *I vitelloni* (1953) utilizzando i descrittori singoli, metodo *10-fold cross-validation.*

I classificatori addestrati con i descrittori di movimento A_M e di intensità del colore A_σ forniscono i risultati peggiori in quanto risentono del non bilanciamento tra primi piani e campi lunghi; il descrittore α di apertura angolare delle linee prospettiche fornisce prestazioni migliori, ma presenta comunque

una bassa accuratezza nella classificazione dei campi lunghi. Per quanto riguarda l'analisi delle proprietà spettrali, le prestazioni sono in media poco più alte rispetto a quelle dei tre descrittori appena visti.

Titolo	Anno	Fotogrammi	Accuratezza (%)
Lo sceicco bianco	1952	4278	72,54
I vitelloni	1953	4483	training
La strada	1954	4261	73,80
Otto e mezzo	1963	6915	training
Fellini Satyricon	1969	6171	training
Roma	1972	4420	57,49
Amarcord	1973	4211	61,90
Il Casanova di Fellini	1976	6689	training
E la nave va	1983	4453	58,12
La voce della luna	1990	4230	55,99

Tabella 3.14 - Stima della SSD su film interi. Il *training set* è composto da quattro film scelti casualmente. Il test viene condotto sugli altri sei.

3.2.4 Addestramento e test su gruppi di film

L'ultima analisi effettuata mira a stimare la distribuzione delle inquadrature di un intero film appartenente all'insieme scelto. L'obiettivo in questo caso è di verificare la capacità del sistema di generare una SSD affidabile per un film totalmente sconosciuto, utilizzando i dati provenienti da altri film per l'addestramento.

Per questo tipo di analisi sono stati scelti casualmente quattro film per generare il *training set* (estraendo un campione bilanciato fra le diverse classi) e i rimanenti sei film sono stati utilizzati per il test. Gli insiemi e i risultati in termini di accuratezza sono indicati in Tabella 3.14.

Le SSD stimate in modo automatico sono poste a confronto con le relative annotazioni manuali in Figura 3.4. Si noti che soprattutto per i film in bianco e nero (*Lo sceicco bianco* e *La strada*) le SSD generate automaticamente sono molto simili a quelle delle annotazioni manuali; negli altri casi mostrano delle piccole variazioni. Allenare il sistema con un insieme bilanciato di campioni ha l'importante conseguenza di compensare gli errori (falsi positivi e falsi negativi per ogni classe sono simili nei numeri) e, nonostante l'accuratezza totale sia molto variabile fra i diversi film, riduce di molto l'impatto di questi errori nella stima delle SSD.

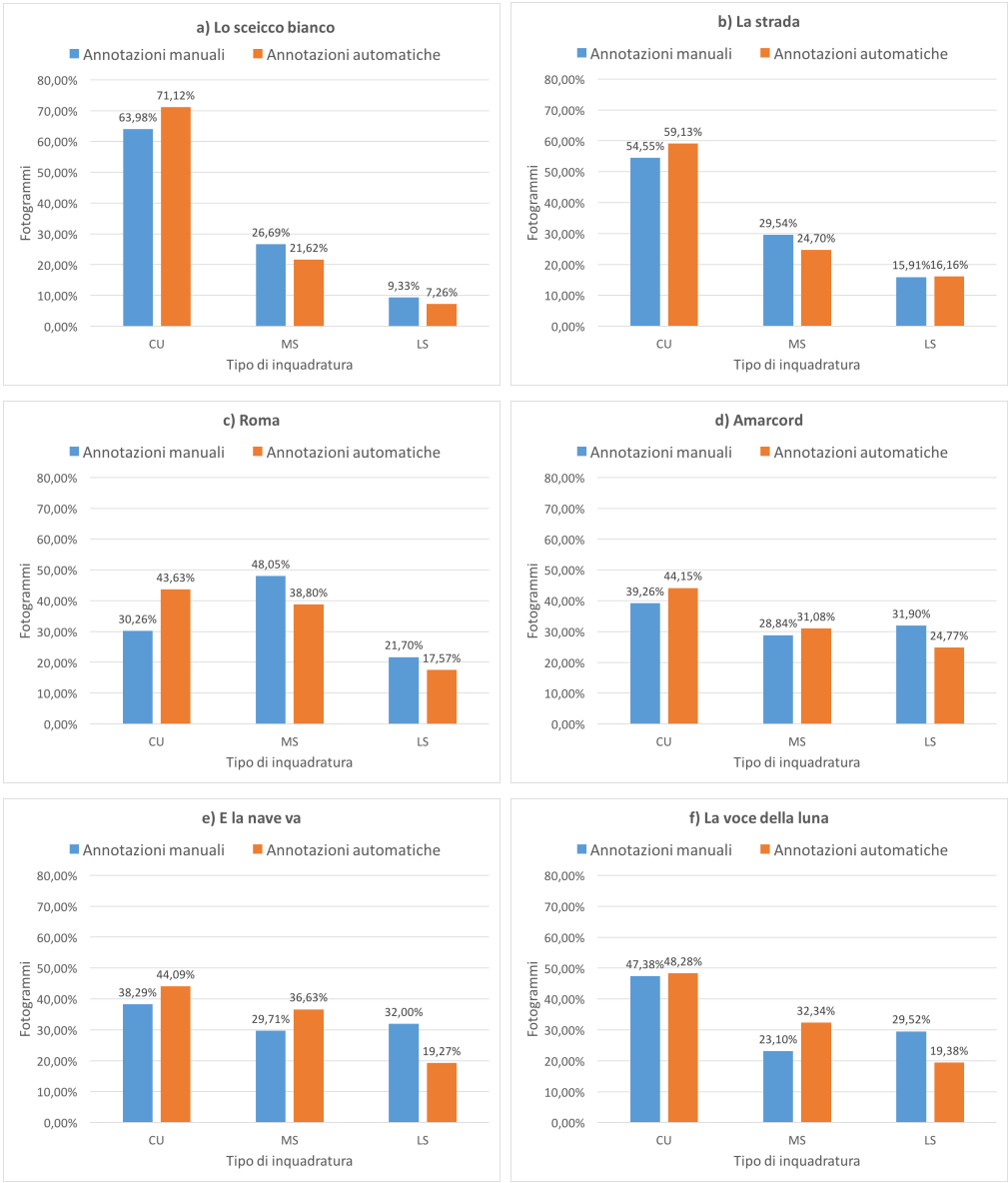


Figura 3.4 - Prestazioni del sistema sulla stima della SSD di film interi. Le annotazioni automatiche sono affiancate a quelle manuali.

(a) (b)

Figura 3.5 - Esempi di possibili errori nelle annotazioni manuali del film *La strada* (1954): in a) osservatori diversi possono classificare diversamente la scena sia come CU che come MS; in b) invece un campo lungo è stato erroneamente annotato a mano come primo piano.

3.3 Analisi degli errori

Da un'analisi condotta sui fotogrammi classificati in modo errato, è emerso che gli errori sono riconducibili a diverse ragioni.

Per prima cosa, bisogna considerare gli errori nelle annotazioni. In numerose situazioni è complicato stabilire con precisione il tipo di inquadratura e anche diversi osservatori umani possono generare annotazioni lievemente diverse fra loro (Figura 3.5(a)). Oltre a questo problema, può capitare che ci siano degli errori nelle annotazioni utilizzate per il *training* e il *testing* del sistema che condizionano inevitabilmente le prestazioni del sistema stesso. Un esempio è fornito in Figura 3.5(b).

Un'altra causa di errore nella classificazione delle inquadrature è la mancata o l'errata individuazione di volti e figure umane nelle immagini. A causa della rilevanza dei descrittori relativi alla presenza umana, quando uno di questi due descrittori sbaglia è probabile che l'errore si ripercuota sulla classificazione. Questo avviene sia in caso di falsi negativi che in caso di falsi positivi, come si vede chiaramente in Figura 3.6.

Altri errori possono essere dovuti alla presenza di oggetti non rilevanti a distanze diverse. In questi casi diventa difficile per il sistema stabilire quale sia il centro dell'attenzione e le proprietà semantiche dell'inquadratura. Un

(a) (b)

Figura 3.6 - Esempi di errori dovuti alla sbagliata individuazione delle figure umane in *La voce della luna* (1990): in a) un CU viene identificato come LS a causa di un errore dell'algoritmo di *body detection*; in b) un falso positivo (box più grande) causa la classificazione di un MS come CU.

(a) (b)

Figura 3.7 - Esempi di errori di classificazione tratti dal film *Otto e mezzo* (1963): in a) il corpo fuori fuoco in primo piano causa il riconoscimento di un CU anziché di un LS; in b) la panoramica verso sinistra della macchina da presa comporta una variazione progressiva del tipo di inquadratura (una persona sta entrando da sinistra).

(a) (b)

Figura 3.8 - Esempi di fotogrammi estratti da *Il Casanova* (1976) per cui è difficile stabilire il tipo di inquadratura anche da un umano.

errore di questo tipo è riportato in Figura 3.7(a). Anche i movimenti della macchina da presa possono indurre errori di classificazione: questi movimenti causano variazioni del tipo di inquadratura all'interno della stessa scena e, soprattutto quando la transizione è lenta, il sistema potrebbe non classificare i fotogrammi in modo accurato durante la transizione, come ad esempio in Figura 3.7(b).

Infine, ricordiamo che i film d'autore contengono spesso esperimenti estetici e inquadrature non tradizionali dall'alto contenuto simbolico che sono in genere molto difficili da classificare anche per uno spettatore umano. Esempi di questo tipo sono riportati in Figura 3.8.

Conclusione

In questo lavoro abbiamo indagato i metodi disponibili allo stato dell'arte per la classificazione delle inquadrature nei prodotti audiovisivi e abbiamo verificato la validità del sistema proposto in [2] conducendo dei test su una parte della filmografia di Federico Fellini. Come si è visto, i risultati ottenuti sono compatibili con quelli riportati dagli autori riguardanti l'analisi della filmografia di Michelangelo Antonioni.

Per quanto riguarda gli sviluppi futuri di questa applicazione, la naturale estensione del lavoro proposto è diretta verso una classificazione automatica più fine (capace ad esempio di distinguere anche primissimi piani, mezzi primi piani o campi lunghissimi) e che comprenda tipologie di inquadratura meno tradizionali come quelle *Over-the-shoulder* e i *Foreground shot*. Data l'elevata modularità del sistema, è anche possibile sperimentare l'impiego di nuove tecniche per l'estrazione dei vari descrittori o utilizzare algoritmi di classificazione differenti.

Non è ancora possibile determinare fino a che punto la SSD possa essere considerata il segno distintivo di un film o di un autore particolari, in quanto l'insieme di film su cui è stato condotto questo tipo di analisi è ancora molto limitato. Tuttavia, in accordo con lo studio sviluppato in [1], la sensazione è che la distribuzione delle inquadrature non solo giochi un ruolo importante per le reazioni che un film provoca nello spettatore, ma possa anche diventare una sorta di "firma" del film stesso. Per verificare queste regolarità nella filmografia di un autore o di un gruppo di film, l'analisi verrà estesa a film di altri autori e di altri periodi. Uno sviluppo interessante potrebbe nascere dalla valutazione sistematica delle SSD su un'ampia collezione di film (tratti per esempio dall'Internet Movie Database [18]) utilizzando metodi di analisi

per grandi quantità di dati.

Bibliografia

[1] A. B. Kovács, "Shot scale distribution: An authorial fingerprint or a cognitive pattern?", *Projections*, vol. 8, n. 2, 2014. DOI: 0.3167/proj.2014.080204.

[2] S. Benini, M. Svanera, N. Adami, R. Leonardi e A. B. Kovács, "Shot scale distribution in art films", *Multimedia Tools and Applications*, pp. 1–29, 2016, ISSN: 1573-7721. DOI: 10.1007/s11042-016-3339-9.

[3] D. Arijon, *Grammar of the Film Language*. Silman-James Press, 1991.

[4] I. Cherif, V. Solachidis e I. Pitas, "Shot type identification of movie content", in *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*, feb. 2007, pp. 1–4.

[5] I. Tsingalis, N. Vretos, N. Nikolaidis e I. Pitas, "Svm-based shot type classification of movie content", in *Proceedings of 9th Mediterranean Electro technical Conference, Instanbul, Turkey, 16-18 October*, 2012, pp. 104–107.

[6] H. L. Wang e L. F. Cheong, "Taxonomy of directing semantics for film shot classification", *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, n. 10, pp. 1529–1542, ott. 2009, ISSN: 1051-8215. DOI: 10.1109/TCSVT.2009.2022705.

[7] S. Bhattacharya, R. Mehran, R. Sukthankar e M. Shah, "Classification of cinematographic shots using lie algebra and its application to complex event recognition", *Multimedia, IEEE Transactions on*, vol. 16, n. 3, pp. 686–696, apr. 2014, ISSN: 1520-9210. DOI: 10.1109/TMM.2014.2300833.

[8] M. Xu, J. Wang, M. A. Hasan, X. He, C. Xu, H. Lu e J. Jin, "Using context saliency for movie shot classification", in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 3653–3656.

[9] L. Xie, S. F. Chang, A. Divakaran e H. Sun, "Structure analysis of soccer video with hidden markov models", in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 4, mag. 2002. DOI: 10.1109/ICASSP.2002.5745558.

[10] A. Torralba e A. Oliva, "Depth estimation from image structure", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, n. 9, pp. 1226–1238, set. 2002, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2002.1033214.

[11] T. Nagai, T. Naruse, M. Ikehara e A. Kurematsu, "Hmm-based surface reconstruction from single images", in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 2, 2002, pp. 561–564.

[12] M. J. Brooks e B. K. Horn, "Shape and source from shading", in *International Joint Conference on Artificial Intelligence*, 1985.

[13] B. Super e A. Bovik, "Shape from texture using local spectral moments", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, n. 4, pp. 333–343, 1995.

[14] H. Barrow e J. Tenenbaum, "Interpreting line drawings as three-dimensional surfaces", in *Artifical intelligence*, vol. 17, 1981, pp. 75–116.

[15] I. Shimshoni, Y. Moses e M. Lindenbaum, "Shape reconstruction of 3d bilaterally symmetric surfaces", in *International Journal of Computer Vision*, vol. 2, 2000, pp. 1–15.

[16] S. E. Palmer, *Vision Science: Photons to phenomenology.* MIT Press, Cambridge, MA, 1999.

[17] D. Hoiem, "Seeing the world behind the image: Spatial layout for 3d scene understanding", tesi di dott., Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, ago. 2007.

[18] *Internet movie database (imdb)*. indirizzo: `http://www.imdb.com/`.

[19] *Wikipedia*. indirizzo: `http://en.wikipedia.org/wiki/Art_film`.

[20] V. Cantoni, L. Lombardi, M. Porta e U. Vallone, "Qualitative estimation of depth in monocular vision", in *4th International Workshop on Visual Form*, Springer, cur., 2001, pp. 135–144.

[21] T. Kurita, N. Otsu e N. Abdelmalek, "Maximum likelihood thresholding based on population mixture models", *Pattern Recognition*, vol. 25, n. 10, pp. 1231–1240, 1992, ISSN: 0031-3203. DOI: `10.1016/0031-3203(92)90024-D`.

[22] O. Barnich e M. V. Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequence", *IEEE Transactions on Image Processing*, vol. 20, n. 6, pp. 1709–1724, 2011. DOI: `10.1109/TIP.2010.2101613`.

[23] A. McIvor, "Background subtraction techniques", in *Proc. of Image and Vision Computing*, 2000.

[24] A. Matessi e L. Lombardi, "Vanishing point detection in the hough transform space", in *Euro-Par99 Parallel Processing*, Springer, 1999, pp. 987–994.

[25] X. Zhu e D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild", in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2879–2886.

[26] P. F. Felzenszwalb, R. B. Girshick, D. McAllester e D. Ramanan, "Object detection with discriminatively trained part-based models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, n. 9, pp. 1627–1645, 2010. DOI: `10.1109/TPAMI.2009.167`.

[27] R. Fisher, "The use of multiple measurements in taxonomic problems", *Annals of Eugenics*, vol. 7, pp. 179–188, 1936. DOI: `10.1111/j.1469-1809.1936.tb02137.x`.

[28] T. Anderson, *An Introduction to Multivariate Statistical Analysis*. Wiley, 1958.

[29] Indirizzo: http://en.wikipedia.org/wiki/K-d_tree.

[30] Indirizzo: https://www.cs.swarthmore.edu/~meeden/cs63/f05/lab05.html.

[31] L. Breiman, "Random forests", *Machine Learning*, vol. 45, n. 1, pp. 5–32, 2001, ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

[32] Indirizzo: http://file.scirp.org/html/6-9101686_31887.htm.