La Computer Vision per l'annotazione automatica di documenti audio

Lauro Snidaro
Laboratorio AVIRES
Dipartimento di Matematica e Informatica
Università degli Studi di Udine

Sergio Canazza
Laboratorio AVIRES
Dipartimento di Scienze Storiche e Documentarie
Università degli Studi di Udine

SOMMARIO
Durante il processo di conservazione attiva dei documenti audio è importante non trascurare l'informazione contestuale presente nel supporto sonoro. In particolare, le riprese video del trasferimento A/D del segnale audio (disco fonografico in rotazione, nastro magnetico in scorrimento, ecc.) possono essere un mezzo per memorizzare le informazioni sullo stato del supporto. Nel caso di archivi audio di rilevanti dimensioni, è economicamente improponibile estrarre manualmente le informazioni di interesse dai filmati. Viene presentato, in questa sede, un sistema in grado di individuare automaticamente eventi di interesse presenti nelle registrazioni video di documenti audio.

Parole Chiave
Annotazione automatica; Visione Artificiale; Conservazione attiva; Documenti audio.

INTRODUZIONE

Dalla carta usata nel 1860 (a cui risale *Au Clair de la Lune*, la più antica registrazione audio di cui ci è stato tramandato il supporto, effettuata da Édouard-Léon Scott de Martinville mediante fonoautografo[1]), sino al moderno Blu-ray Disc, il campo delle memorie audio costituisce una vera Torre di Babele: in 150 anni sono stati prodotti un numero enorme di supporti (analogici e digitali) incompatibili tra loro. Sono chiare l'urgenza, l'importanza e la complessità della loro conservazione, che può essere articolata in: a) passiva[2] (difesa del supporto dagli agenti ambientali, senza alterarne la struttura) e b) attiva (trasferimento dei dati nei nuovi media). Poiché le memorie audio sono caratterizzate da un'aspettativa di vita relativamente bassa – se confrontata con quella di altri monumenti – la conservazione passiva risulta insufficiente. Inoltre, i beni culturali musicali non hanno la possibilità di vicariare o rigenerare – partendo dal segnale audio – l'informazione perduta: è quindi di fondamentale importanza trasferire nel dominio digitale (oltre al segnale audio) tutta l'informazione contestuale presente nel supporto. In questo senso, molti archivi inseriscono nella copia conservativa le riprese video effettuate durante il trasferimento A/D del segnale audio (scorrimento del nastro, rotazione del disco o del cilindro di cera). Il video, infatti, offre informazioni sulle alterazioni (intenzionali o meno) e sulle corruttele del supporto. Questi dati sono necessari a fini archivistici, per studi musicologici e nelle operazioni di restauro audio [1], [2], [3].

Nel caso di archivi di medie/grosse dimensioni, è economicamente improponibile estrarre manualmente le informazioni di interesse dai filmati. Diventa quindi prezioso poter contare si di un sistema in grado di individuare automaticamente eventi di interesse presenti nelle registrazioni video di documenti audio.
Gli autori hanno utilizzato strumenti sviluppati nell'ambito della visione artificiale per rilevare automaticamente discontinuità presenti nel nastro magnetico. In particolare, sono state impiegate tecniche di *background subtraction* con impostazione automatica della soglia [4], ottimizzata al fine di individuare la presenza di nastro *leader*. In questo modo sono annotati automaticamente gli istanti temporali d'inizio e di fine di ogni trato di nastro magnetico (registrabile), rispetto al nastro di plastica utilizzato durante la fase di montaggio dai compositori e dai tecnici. Nel caso il nastro coprisse solo una percentuale dell'immagine (caso molto comune), è possibile impostare una *regione di interesse* (*Region Of Interest,* ROI), in modo da scartare durante l'elaborazione i dettagli non rilevanti (testina di lettura, sfondo, ecc.). Questo approccio è mutuato dalle tecniche utilizzate per il rilevamento dei cambi di scena nel campo dell'annotazione automatica di sequenze video [5].
A questa tecnica di base sono poi stati accoppiati altri algoritmi al fine di rilevare specifiche alterazioni sul nastro (intenzionali o meno) nelle sotto-regioni selezionate (ossia: in cui il sistema ha rilevato una discontinuità). In questo modo vengono automaticamente annotati gli istanti temporali in cui nel nastro compaiono: a) giunte; b) segni; c) perdite di pasta magnetica. Nel caso in cui al video fosse sincronizzato il segnale audio, ognuna di queste discontinuità può essere facilmente allineata al corrispondente evento sonoro.

Nel caso di videoregistrazioni di dischi fonografici in rotazione, il sistema rileva l'evoluzione temporale della posizione del braccio. Da questa funzione, può essere automaticamente calcolata (conoscendo la velocità di rotazione usata per leggere il disco) la variazione del *pitch* nel segnale audio dovuta alle deformazioni presenti nel supporto

[1] Brevettato il 25 marzo 1857. Utilizzato come strumento di laboratorio per studi di acustica, con funzione simile all'oscilloscopio, era in grado di trascrivere graficamente le onde sonore su un mezzo visibile (vetro annerito o rotolo di carta), ma non c'era modo di riprodurre il suono registrato.

[2] La conservazione passiva si suddivide a sua volta in indiretta – che non comporta il coinvolgimento fisico del supporto sonoro – e diretta, nella quale il supporto viene trattato, senza comunque alterarne struttura e composizione. Nella conservazione passiva indiretta rientrano: la prevenzione ambientale (che si esplica attraverso il controllo dei parametri ambientali che sono, in ordine decrescente di pericolosità per i nastri magnetici: umidità relativa, temperatura, inquinamento, luce), la formazione del personale addetto alla conservazione, l'educazione dell'utente. La conservazione passiva diretta comprende gli interventi di: realizzazione di custodie di protezione; spolveratura delle raccolte; disinfestazione degli archivi con gas inerti; periodico svolgimento e riavvolgimento dei nastri magnetici.

fonografico rispetto al suo piano (dischi cosiddetti *imbarcati*), oppure a difetti di bilanciamento del piatto. Questo dato è di estrema utilità nel caso si voglia procedere, in fase di restauro audio, a riconoscere le oscillazioni del *pitch* dovute a: a) difetti del sistema di registrazione; b) deterioramento del supporto; c) imperfezioni del sistema di lettura.
L'approccio sperimentato si basa sull'analisi dello spostamento delle *features* di Lucas-Kanade rilevate sul braccio del giradischi. La tecnica sviluppata consiste in primo luogo nel *clustering* delle *features* non statiche all'interno di una ROI prefissata; viene quindi valutata la variazione di inclinazione del *cluster* nel tempo determinando quindi la variazione di *pitch*.

DISCONTINUITA' PRESENTI SU NASTRI MAGNETICI

La ripresa dello svolgimento di un nastro, durante il trasferimento dei dati audio in esso contenuti su supporto digitale, permette di raccogliere importanti informazioni utili allo studio del segnale registrato.
Una sequenza video costituisce una ricca fonte di informazioni sullo stato del nastro. In particolare, è possibile evincere alterazioni quali corruttele del supporto o interventi di editing (giunte, segni). Queste informazioni rappresentano preziosi metadati con cui annotare la copia riversata su supporto digitale. Essi possono essere infatti associati all'informazione acustica fornendo importanti elementi per un'analisi della storia del documento audio originale e per il suo restauro.

L'estrazione manuale di metadati mediante ispezione di sequenze video da parte di un operatore (annotazione manuale) non è concepibile per archivi di medie/grosse dimensioni [6]. In questo lavoro vengono quindi proposte delle tecniche di visione artificiale in grado di elaborare automaticamente le sequenze video ottenute dalle riprese di documenti audio su nastro durante la loro riproduzione. Queste tecniche sono in grado di rilevare le discontinuità presenti sul nastro e di riconoscere elementi notevoli quali giunte.

Negli esperimenti effettuati, la tecnica di base impiegata per rilevare discontinuità in una sequenza video è la nota *background subtraction*: il fotogramma corrente $\boldsymbol{I}_t$ all'istante t viene confrontato con uno di riferimento $\boldsymbol{I}_{bck}$ (background), acquisito in un istante temporale precedente. Il confronto, effettuato mediante differenza in valore assoluto fra matrici, permette di ottenere l'immagine delle differenze $\boldsymbol{D}_t=|\boldsymbol{I}_t - \boldsymbol{I}_{bck}|$. All'immagine delle differenze $\boldsymbol{D}_t$ viene applicata una soglia *th* ottenendo quindi un'immagine binaria $\boldsymbol{B}_t$. Un'immagine binaria può rappresentare solo due colori, convenzionalmente il bianco e il nero. Con il nero vengono in genere indicate le regioni dell'immagine che non hanno subito cambiamenti, mentre in bianco vengono indicate le differenze sostanziali fra le due immagini oggetto del confronto, quelle cioè che hanno superato la soglia *th*. Negli esperimenti qui proposti è stato impiegato l'algoritmo di *sogliatura automatica* basato sui numeri di Eulero i cui dettagli sono riportati in [4]. Questo algoritmo calcola automaticamente per ogni frame la soglia *th* ottimale in base ad un criterio legato alla connettività delle componenti connesse presenti in $\boldsymbol{B}_t$. E' stato inoltre applicato un passo di filtraggio come ulteriore elaborazione. In particolare, è stato impiegato un semplice *voting* per rimuovere piccole componenti connesse spurie, dovute a rumore o a improvvise variazioni di luminosità locale, nella matrice binaria $\boldsymbol{B}_t$ al fine di migliorarne la qualità.

In Figura 1e in Figura 2 sono visibili alcuni fotogrammi estratti da sequenze video diverse su cui sono stati condotti gli esperimenti. Nella prima colonna sono stati riportati i frame sorgente, mentre nella seconda sono state riportate le immagini binarie risultato delle procedure di *change detection* e di *sogliatura*. Per entrambe le sequenze, il fotogramma di background era stato acquisito precedentemente ai fotogrammi illustrati.

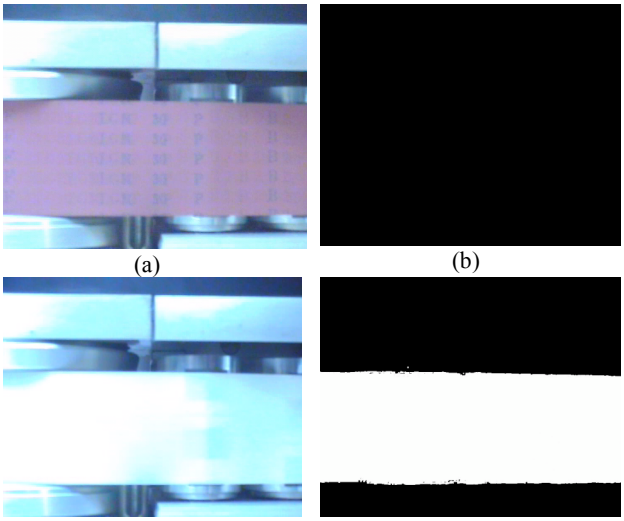

Figura 1. Nella prima colonna sono presenti due immagini estratte dal video di un nastro in fase di riproduzione. Nella seconda colonna sono visibili le corrispondenti elaborazioni. (a) Dorso del nastro magnetico e (b) nessuna anomalia rilevata. Il nastro *header* in (c) viene rilevato come discontinuità (d).

La Figura 1 evidenzia come la tecnica della *change detection* permetta di rilevare discontinuità nella composizione del nastro. La Figura 1(b) è completamente nera poiché non è stata rilevata nessuna differenza significativa tra il fotogramma corrente (a) e quello di background. In (d) viene invece rilevata una differenza consistente (regione bianca) fra l'immagine di riferimento e quella corrente (c). In questo caso, è evidente come la sezione *header* del nastro viene riconosciuta con precisione.

Le procedure sopra descritte (si consulti [4] per ulteriori dettagli realizzativi) forniscono quindi un mezzo efficace per rilevare variazioni nella composizione del nastro. Il conteggio dei pixel bianchi nelle immagini binarie e una soglia fissata a priori sulla percentuale di pixel cambiati rispetto alla *regione di interesse* (*Region Of Interest*, ROI) permettono di decidere se sì è in presenza di una discontinuità o meno. La ROI può essere impostata per focalizzare l'attenzione dei vari algoritmi solo su una sottoregione dell'immagine. Come si può vedere nei frame sorgente della Figura 1, il nastro occupa circa il 50% dell'immagine; mentre altri dettagli come le testine del lettore non sono rilevanti ai fini dell'elaborazione e dovrebbero essere eliminati impostando la ROI sulla regione corrispondente al solo nastro.
L'approccio appena descritto è molto simile alle tecniche di *scene cut detection* per l'annotazione automatica di filmati televisivi o cinematografici [5], [6].

La Figura 2 illustra invece come altri tipi di informazioni possano essere estratte mediante l'analisi delle riprese di nastri in svolgimento. I passi di elaborazione di base sono gli stessi dell'esperimento precedente, ma in questo caso sono richiesti dei procedimenti aggiuntivi per rilevare giunte o discontinuità specifiche.

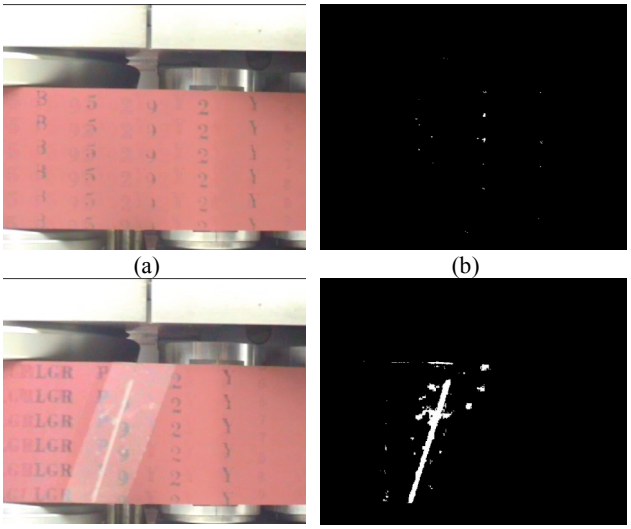

Figura 2. Discontinuità rilevate su un nastro in fase di riproduzione. (a) Dorso del nastro magnetico e (b) nessuna anomalia rilevata. La giunta visibile in (c) viene rilevata come discontinuità dal sistema (d).

Il frame (b) è quasi completamente nero ad indicare che non vi sono cambiamenti significativi fra il frame corrente (a) e l'immagine di riferimento. Le piccole componenti connesse osservabili in (b) rappresentano delle piccole variazioni dovute allo scorrimento dei caratteri alfanumerici visibili in (a). Questi piccoli cambiamenti (numero di pixel bianchi) non superano la soglia di attenzione impostata a priori dall'operatore e quindi non rappresentano un'anomalia del nastro. Il fotogramma (d) evidenzia come la giunta venga rilevata dalla change detection. In questo caso, il numero di pixel cambiati passa la soglia di attenzione. Tuttavia, al fine di discriminare il tipo di anomalia è necessario un passo ulteriore di elaborazione. L'individuazione del segmento in (d) corrispondente alla giunta può essere effettuata mediante l'applicazione della trasformata di Hough all'immagine binaria [7]. Queste informazioni – opportunamente allineate col segnale audio – possono essere annotate automaticamente dal sistema direttamente nella copia conservativa.

E' utile notare che le tecniche impiegate, in particolare la *sogliatura automatica* basata sui numeri di Eulero, hanno consentito di analizzare sequenze video acquisite senza particolari accorgimenti volti a creare delle condizioni normalizzate. In particolare, si può osservare come i fotogrammi in Figura 1 e in Figura 2 siano stati acquisiti in condizioni di illuminazione sensibilmente differenti. Questo facilita di molto le videoregistrazioni stesse che possono essere effettuate dagli archivi senza vincoli particolari sulle condizioni di illuminazione, mediante hardware *entry level* e senza l'utilizzo di personale addestrato a riprese video professionali.

DEFORMAZIONI DEI SUPPORTI FONOGRAFICI

Le deformazioni dei supporti fonografici (dischi *imbarcati* o sbilanciati) causano una sensibile variazione del *pitch* del segnale audio (la Figura 3 mostra un esempio di disco afflitto da un'importante deformazione). Questo è dovuto al movimento ondulatorio a cui è soggetto il braccio del giradischi (e quindi la puntina).

Figura 3. Disco fonografico afflitto da un'importante deformazione rispetto al suo piano. La corruttela è stata causata da un'escursione della temperatura (il disco è stato esposto al sole per 4 ore, all'interno di un'autovetura in sosta, a una temperatura di oltre 50°C).

Poiché le caratteristiche del moto oscillatorio possono essere messe in relazione con la variazione del *pitch*, esse costituiscono importanti metadati utili al restauro del segnale audio. Si propone in questa sede di utilizzare tecniche di visione artificiale per analizzare e annotare automaticamente le videoregistrazioni di dischi fonografici in rotazione.

In questo esperimento è stato impiegato un approccio diverso rispetto a quello basato sulla *change detection* visto precedentemente. In questo caso è stato utilizzato l'algoritmo di *features tracking* comunemente noto come Lucas-Kanade *tracker* [8]. L'algoritmo individua dei punti notevoli nell'immagine (*features*) che possono essere impiegati per confrontare due frame successivi e valutare gli spostamenti fra uno e l'altro. La tecnica, inizialmente concepita per l'allineamento di immagini, viene qui utilizzata nella sua implementazione come *features tracker*, che è in grado di tenere traccia degli spostamenti delle *features* da un frame al successivo.

Nella Figura 4 sono stati riportati alcuni fotogrammi tratti da una delle sequenze impiegate negli esperimenti. La prima colonna riporta (a) il punto più elevato e (d) quello più basso dell'oscillazione del braccio. Nella seconda colonna si possono osservare le rispettive elaborazioni in cui sono visibili le feature individuate sul braccio del giradischi.

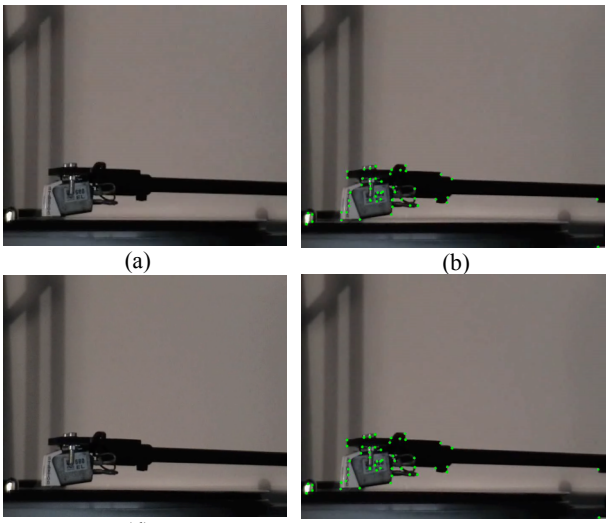

Figura 4. Immagini sorgente e relative elaborazioni tratte da una ripresa video del braccio di un giradischi durante la riproduzione. Il supporto sul piatto è deformato e ciò causa un'oscillazione del braccio. Nelle immagini sono state evidenziate: (a) posizione più bassa del braccio durante l'oscillazione e (d) posizione più alta. Nelle immagini (b),(e) si possono notare le *features* di Lukas-Kanade rilevate sulla testa del braccio.

Durante gli esperimenti il *tracker* di Lucas-Kanade ha correttamente mantenuto traccia delle *features* rilevate nel primo frame delle sequenze video utilizzate. Il *tracker* ha quindi permesso di registrare gli spostamenti delle *features* durante le oscillazioni del braccio del giradischi dovute alla riproduzione di supporti deformati.

Nella Figura 5 è stata riportata l'evoluzione temporale della coordinata *y* di una *feature* localizzata sul braccio. Sull'asse delle ordinate è riportato il numero di *frames*, mentre le ascisse rappresentano la posizione in pixel sul piano immagine. E' chiaramente visibile l'andamento oscillatorio. I fotogrammi (a) e (d) nella Figura 4 distano 29 frame, dato riscontrabile nelle oscillazioni del grafico di Figura 5.

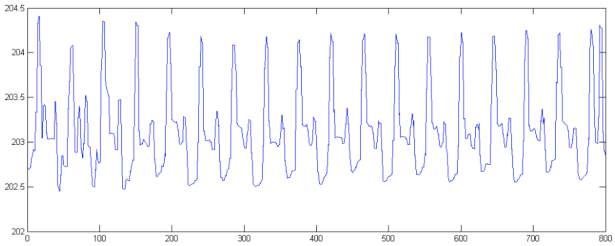

Figura 5. Evoluzione temporale della coordinata y di una feature di Lucas-Kanade posizionata sul braccio. Le evidenti oscillazioni indicano la presenza di un supporto imperfetto.

CONCLUSIONI

Nel caso di archivi sonori di medie/grosse dimensioni, è economicamente improponibile estrarre manualmente le informazioni di interesse dai filmati effettuati per conservare l'informazione contestuale dai supporti fonografici. Diventa quindi prezioso poter contare si di un sistema in grado di annotare automaticamente eventi di interesse presenti nelle registrazioni video dei dischi in rotazione. Questi metadati possono essere vantaggiosamente utilizzati in fase di restauro del segnale audio per riconoscere le alterazioni inserite durante la fase di registrazione/produzione del supporto, quelle dovute al deterioramento del supporto o alle imperfezioni del sistema di lettura.

Le applicazioni sono state sviluppate in C++ e sono in grado operare in tempo reale su segnali video a risoluzione di 320x240 *pixels*, e utilizzando un elaboratore dotato di (singolo) processore a 3 GHz. Le sequenze video possono essere acquisite mediante un *camcorder* a risoluzione PAL e successivamente scalate e compresse in DivX a qualità medio-alta.

Grazie alle basse richieste computazionali e al fatto che non viene richiesto hardware specifico, si ritiene che il sistema possa essere un pratico ausilio per gli archivi di documenti sonori e le audio Digital Libraries, che possono in questo modo annotare automaticamente l'informazione contestuale dei supporti fonografici utilizzando personale con modesto addestramento.

RIFERIMENTI

[1] Orcalli, A. (2006). Orientamenti ai documenti sonori. In Canazza, S. e Casadei Turroni Monti, M. (a cura di*), Ri-mediazione dei documenti sonori*, pp. 15-94, Udine: Forum.

[2] Canazza, S. (2007). Note sulla conservazione attiva dei documenti sonori su disco. In Atti del Convegno annuale del Laboratorio per la Divulgazione Musicale (Ladimus). *Il suono riprodotto: storia, tecnica e cultura di una rivoluzione del Novecento*, Torino: EDT.

[3] Canazza, S. (2006). Conservazione attiva e restauro audio dei 78 giri. Un caso di studio: Eternamente, In Canazza, S. e Casadei Turroni Monti, M. (a cura di*), Ri-mediazione dei documenti sonori*, pp. 695-715, Udine: Forum.

[4] L. Snidaro and G.L. Foresti, "Real-time thresholding with Euler numbers", *Pattern Recognition Letters*, Vol. 24, n. 9-10, pp. 1533-1544, June 2003.

[5] Y. Liu, D. Zhang, G. Lu, and W.Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, 2007.

[6] A. Del Bimbo, Visual information retrieval, Morgan Kaufmann, San Francisco, CA, 1999.

[7] N. Aggarwal and W.C. Karl, "Line detection in images through regularized hough transform" *IEEE Transactions on Image Processing*, Vol.15, n°3, March 2006.

[8] J. Shi and C. Tomasi, "Good Features to Track", IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WS, June 1994, pages 593-600.