マルチメディア通信と分散処理 99−11
（2000. 9. 22）

音情報と画像情報を用いた動画高速閲覧のための一考察

青柳滋己　高田敏弘　佐藤孝治　菅原俊治
NTT 未来ねっと研究所

CATV や BS,CS などの一般放送の多チャンネル化が進み、また高速なネットワークの広がりにより、TV や PC において多くの映像情報をより簡単に入手できる環境が急速に広がりつつある。このような環境下では氾濫している多くの映像情報から自分の目的にあった情報を短時間で得ることは今後ますます困難になっていくと予想される。映像情報を短時間で見る方法としては一般に早送りが使われるが、ビデオ等で使われる早送りは映像が一定速度で再生され、音声は再生されないか、同期せずに一部だけ再生されるものがほとんどであり、音声と画像の同期がとれてないためわかりにくく、画面を注視しないと内容を理解できないという問題があった。本稿では音声情報を主体にし、画像情報も用いた動画高速閲覧のための仕組みについて述べる。

A Study of Video Skimming based on Audio and Image Recognition

Shigemi Aoyagi　Toshihiro Takada　Koji Sato　Toshiharu Sugawara
NTT Network Innovation Laboratories

Progress of computer network technologies and multiple channels of satellite broadcasting produce a flood of movie information, and users find it difficult to acquire information they want. To shorten the time of viewing movie information, fast-forwarding is the most popular technique. However, current fast-forwarding focuses only on increasing the speed of displaying video images and audio part is not played in many systems. In this paper, we propose a video skimming mechanism which is based on information of audio part and scene changes in video part.

1　はじめに

高速なネットワークの普及により居ながらにして世界中の情報にアクセスすることができる環境が容易に入手できるようになり、情報内容も従来のテキストや静止画に加え、音声や動画などのより情報量の多いものへ変遷しつつある。また、CATVやBS,CSデジタル放送といったテレビ放送も多チャンネル化が進み、ユーザが多チャンネル多ジャンルの映像の中から見たいものを自由に選ぶことが可能な時代へと移りつつある。このように映像情報がますます氾濫していく中、ユーザはそれらの中から効率良く自分の欲する情報を取捨選択していく必要がある。しかしながら映像情報は、テキスト情報における速読のように同じ情報量を短時間で得ることは難しい。現在、映像を短時間で見る場合に最もよく使われるのが早送りである。しかし、現状の早送りは映像を通常の2倍、3倍といった一定の速度で滑らかに再生することにのみ主眼がおかれており、音声情報は全く再生されないか、あるいはところどころを再生するだけである。したがって、ユーザは早送り時には高速に再生される画像部分から目が離せず、しかも音声パートに含まれる情報は十分に再生されず入手できないことが多い。通常の映像情報では音声は効果音等の映像を補助する付加的な情報だけでなく、画像の細かい説明がなされたり、あるいは画像を補助として用い、重要なことは音声で話されている場合も多いため、音声情報は早送りの間でも再生されるのが望ましい。また、画像と音声は、独立に再生したのでは関連付けが難しくなるため、なるべく同期して再生するのが望ましい。

本稿では音の中の音声情報と画像のカット点検出の情報を用い、映像と音声の同期をとりながら再生を行う動画情報閲覧時間短縮法について述べる。

2　市販システムおよび関連研究

動画の高速再生が最も使われているのは市販ビデオデッキである。通常、早送りでは映像のみが2倍～7倍速程度のスピードで再生され、音声は再生されないものが多い。しかし、一部の会社の製品には映像だけでなく音声も再生するものがある。ビクター社のタイムスキャン、サンヨー社の時短ビデオなどが有名である。

ビクター社のタイムスキャンでは、映像は7倍速度固定で再生され、音声は飛び飛びに数秒間ずつ再生される。早送りだけでなく巻き戻し時にも音声を再生する点に特徴がある。再生される音声は完全に飛び飛びのため、会話場面等では会話の内容を把握するのは困難であり、また再生音声は映像と全く同期しない。最近の新製品では1.5倍速度で音声と映像の同期を取るモードがある。サンヨー社の時短ビデオはビクター社のものよりもう少し高度である。映像はビクター社の製品と同じく一定倍率固定で再生されるが、音声は無音部分をカットし速度等を変化させてできるだけ漏れのないように再生させる。ただし、どうしても間に合わない部分はカットされる。また、音声は映像とは同期せず、遅れて再生される。

動画の研究としては映像理解やデータベース構築のための研究は数多く行われてきた。例えば柿沼ら[1]は映像だけでなくシナリオがあることを前提に、ドラマ映像を対象にしたデータベースの構築をめざしたものである。映像の特徴抽出を行う手段として、映像全体の構造を認識するためのカット点の検出やフェイドイン、ワイプといった特殊効果を検出する方法も検討されている[4]。これらの研究は映像をデータベース化することで検索等を容易にする目的のものが多い。

音の情報を積極的に利用し映像のインデクシングを行う研究には[2]がある。音声パート

の認識や音楽のスペクトル特徴を利用して、音楽、男性の声、女性の声を分別している。

いずれの研究も映像に対して、検索のためのインデクシング等を行うための研究である。効果的なインデクシングが行われれば膨大な映像情報から目的の映像を抜き出すことができるが、本研究の閲覧時間短縮はそれらとは独立に用いることができる。例えば、そのインデクシングによって入手した動画データに対してこれから述べる閲覧法が使えるのである。

3 音声を重視した高速再生

従来の動画早送りや高速再生では次のような問題点があった。

1. 映像を見続けないと内容をのがしてしまう。特に高速で再生されているため、一瞬でも目を離すとまったく別の画面になってしまうので注視しなければならず、気が抜けない。

2. 音声が全く再生されないか、一部しか再生されないことが多く、音声部分に含まれる情報を得ることが困難である。

そこで、本稿では音声を主体にした動画閲覧法を提案する。音声を主体にした閲覧法では、まず音声のない映像部分は重要でないと考え、その部分は再生しないことにする。実際に使われる映像では、音声を含まない部分は短く、音声を含む部分全部をそのまま再生したのでは閲覧時間の短縮にはほとんどならないことが考えられる。そこで、次に音声を含む部分の高速化をする。ただし、そのまま高速に再生したのでは音声ピッチがあがってしまうので、後に述べる話速変換を用いてピッチがあがらないようにする。映像部分については、音声と同期して再生を行う。音声のない映像の場合にはこのままではうまくいかない。そこで、画像部分から場面の切り替わるカット点の検出を行い、カット点以降の画像をしばらく再生させることにした。映像内容が変わる場面では変わった後の映像が再生され、ほとんど変化のない部分は飛ばされることになる。

この方式を用いると、音声部分の情報がほぼ再生されるため、画面をずっと見続けなくても音声からも十分に必要な情報を得ることができる。また、画面と音声が同期して再生されるため、自然な再生が可能となり、上で述べた従来の高速再生の問題点を解決できる。

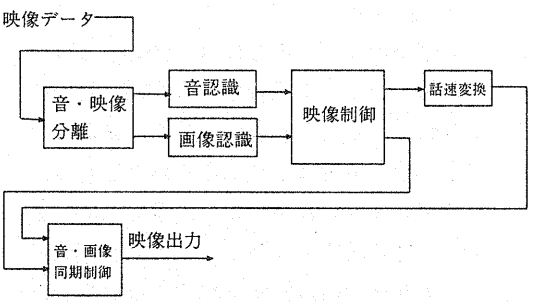

図 1: 処理の流れ

図 1は本方式を実現するためのシステムの処理の流れを表したものである。個々の処理は以下のようになる。

1. 動的特徴量を用いた音声区間の認識
   動画中の音データを一定周期のフレームごとにケプストラム分析を行い、その各ケプストラムの各次において求めた係数の2乗和であるΔケプストラムの値は、スペクトルのゆるやかな動きに比例する。音声はこのゆるやかな動きが生じるのでΔケプストラムの値は大きくなり、定常的な雑音などはスペクトル変化がないため、Δケプストラムの値は小さくなる。

このΔケプストラムの値を調べることにより、無音区間や定常雑音の区間と音声区間を判別することが可能である。ただし、楽器等で演奏された音楽も音声と同じくスペクトルの緩やかな変化がありΔケプストラムの値は大きくなるため、音声と音楽を区別することはできない。

この動的特徴量の計算はリアルタイム計算が可能であるが、この計算によって求めた特徴量を用いてその区間の音声を再生するかしないかを決定する必要があるので、あらかじめ計算しておき、再生にはその計算した値を用いる。

2. 音の高速化

放送映像などの場合には、無音部分はほとんどなく、大抵音楽や音声が切目なく流れていることが多い。また、講演等でも上手な話者になればなるほど必要以上の間を入れずに話すことが多いため、音声以外の区間を削除する、という方式では十分な時間短縮が実現できない場合が多い。音声区間については音の再生速度をあげて再生させれば時間短縮がはかれる。ただし、データをそのまま高速再生させたのでは音のピッチが変化してしまう。そこで音のピッチが変化しない話速変換手法を用いて高速化した。　話速変換は、まず波形データに窓関数をかけて微細部分に区切り、その微細部分を少しずつずらして足し合わせることで実現している。使用した窓関数はGaussian関数、窓長は150ms、窓幅は50msである。

この話速変換のアルゴリズムは処理が簡単であるので、再生時にリアルタイム処理を行っており、変換の倍率は再生中に自由に変更可能である。

3. 画像のカット点検出

Δケプストラムによる音区間認識だけでは音がない区間については完全にスキップされてしまう。そこで、映像の切り替わるカット点の検出を行い、カット点以降の映像についても再生することにした。これにより無音区間の映像もあるていど見逃すことなく再生可能である。

カット点検出のアルゴリズムとしては[?]の分割χ^2検定法を用いた。

分割χ^2検定法では、まず1枚の画像を4x4の16矩形領域に分割し、それぞれの矩形において64色種のヒストグラムを調べる。比較する2枚の画像をf_1, f_2とし、矩形rの色iの色濃度を$H(f_1, r, i)$で表すと、各矩形におけるχ^2検定の値は

$$\sum_{i=0}^{63} \frac{(H(f_1,r,i) - H(f_2,r,i))^2}{H(f_1,r,i)}$$

となる。分割χ^2検定では、$r = 0..15$の全16個の値のうち小さい方の8個の和を評価値とする。

この方式はかなり計算時間がかかるため、あらかじめ計算しておく。

4. 映像のマーク機能

この機能は時間短縮に直接関係ないが、映像のある区間をマークし、いつでも呼び出せるという機能を設けた。この区間は任意の個数登録できる。登録はキーボード登録したい最初のフレームと最後のフレームで特定のキーを押すことにより行われる。また、停止中・再生中に特定のキーを押すことによりマークした映像部分を呼び出すことが可能である。

4　実験

本方式を用いることによりどのくらい時間短縮が図れるかを実験してみた。3つの映像に対して本方式の再生時間を計測した。結果を表1に示す。映像1は料理番組、映像2が講演、映像3が複数人登場バラエティ番組である。いずれの映像においても動的尺度のパラメータは人手により最適にあわせており、話速変換は標準速度の1.2倍に設定した。

実験結果では、オリジナルの映像のおよそ60%前後の時間で再生が可能となっている。この短縮率は元の映像においてどの程度音声でない区間があるかに大きく依存するものの、十分実用的であると考える。また、実験の話速変換速度はどんな話者でも十分聞き取れる程度ということで1.2倍速に設定したが、データによってはもっと速度をあげても聞き取れる場合もあり、その際にはより短縮化が可能である。

5　考察および今後の予定

現在のテレビ放送等で使われる映像データのうち音声が入らない映像は少ない。音声部分が少ない映像についても、重要な場面では音声による解説が入る場合が多く、本稿で述べた閲覧時間短縮法は有効に働くと考えている。また、本短縮法では音声が絶え間なく入っているような映像の場合でも話速変換を用いてピッチを変化させることなく再生速度を速めて短縮をはかっている。話速変換による高速化は人間の聞き取り精度の個人差や再生する音データの内容にも依存するが1.5倍程度が限界であり、それ以上高速化しても聞き取れなくなる。音声部分の多い映像では音声部分抜き出しのみによる短縮効果は少なく、また話速変換のみによる短縮効果も少ないが、その両方を用いることでかなりの短縮が可能となり、実験ではオリジナルの60%前後にまで短縮化が図れた。また、音声データのない映像においても、画像のカット点直後の画像を表示することにより画像の粗筋的なものはわかるようにして対応しており、本短縮法は有効な手法であると考えている。

一般家庭におけるテレビ視聴では、常に画面を注視しているのではなく、音を聞きながら興味のある音や音声が聞こえたときに画面を見るという場合もあり得る。しかしながら、従来の早送り法等ではそれまで画面を見ていない状態で画面を見ても画像と音の同期がとれていないため、画面を見ても何のシーンを再生しているのかしばらく判別できないことがある。しかし、本短縮法では音と画像は常に同期して再生されているため、そのような場合でも映像内容がすぐに理解しやすいという特徴がある。

本短縮法は講演や解説など音データが重要な映像に対してうまく機能する。しかし、本短縮法がうまく機能しない種類のデータもある。まず、音のない間の部分まで計算されて作られた映像、映画等がある。この種類の映像については、制作者が意図したように再生されるべきものであるため、本短縮法は対象外である。また、音声が画像に対して遅れがちになるものでは、ユーザが見たいシーンが再生されない場合がある。例えば、一瞬で勝敗を決するスポーツなどの場合、勝敗が決める瞬間の前には余計な音声の解説がなくなり、決した後に解説が入るということが多いため、本方式の再生法では肝心なシーンの映像が再生されないことになってしまう。そういう種類の映像については、音声部分だけでなく音声の少し前から再生する、という方法で対処可能と思われる。これについては現在プレーヤに機能を付加して実験を試みている。

本稿では映像中の画像情報と音情報だけを

	オリジナル (s) $[a]$	再生時間 (s) $[b]$	短縮率 (%) $[\frac{b}{a} \times 100]$
映像 1	749	438	58.48
映像 2	1495	982	65.69
映像 3	1199	848	56.72

表 1: 実験結果

用いた閲覧時間短縮法を検討し実装したが、この 2 つの情報だけでこれ以上の短縮化を図るのは難しい。さらに再生時間を短縮するには、要約システムなどのように映像の構造や意味などの情報を用いて重要度の低い部分の画像・音を削除していく必要がある。我々は、現在規格制定中の MPEG-7 の Multimedia Description Scheme(MDS)[5] を使い、映像の意味的情報から重要でない映像部分を削除してさらなる時間短縮を行うための枠組みの検討及びプロトタイプ試作を進めている。

6 まとめ

本稿では音声情報をベースに画像情報も用いた映像閲覧時間短縮の提案を行った。従来の映像主体の映像再生法では音声情報は無視されるか一部のみの再生しか行われないのに対し、音声を重視した本稿の方式では音声情報は十分に再生され、画像のみの映像に対しても短縮化を行うことが可能であり、また音声と映像は常に同期して再生することが可能となった。

謝辞

本研究について、ご支援下さった小柳恵一未来ねっと研究所ネットワークインテリジェンス研究部部長、及びネットワーク情報処理研究グループの各氏に感謝いたします。

参考文献

[1] 柿沼、坂内: “DP マッチングを用いたドラマ映像・音声・シナリオ文書の対応付け手法の一提案,” 電子情報通信学会論文誌, D-II, Vol.J79–D–11, No.5, ppp.747–755, 1996.

[2] Kenichi Minami, Akihito Akutsu, Hiroshi Hamada, and Yoshinobu Tonomura: “Video Handling with Music and Speech Detection,” IEEE Multimedia Magazine, Vol.5, No.3, pp.17–25, 1998.

[3] 水野、高橋、嵯峨山: “スペクトルの動的および静的特徴量を用いた言語音声の検出,” 日本音響学会講演論文集 (秋), 3-2-1(1995).

[4] 田村、池田 (編): “知能情報メディア”、総研出版、 1995.

[5] Peter van Beek, Ana B. Benitez, Joerg Heuer, Jose Martinez, Philippe Salembier, John Smith, TOby Walker: “MPEG-7 Multimedia Description Schemes WD(Version 2.0),” ISO/IEC JTC 1/SC 29/WG 11/N3247, March 2000.