論文の内容の要旨

論文題目：音声に基づく映像インデクシングに関する研究

氏　名　：　小河　誠巳

近年の情報処理分野の発展に伴い，記録装置の大容量化，DVD メディアの普及，BS，CS ディジタル放送，さらに地上波ディジタル放送の開始と映像コンテンツを取り巻く環境は大きな変動の時期にある．特に個人を中心とした映像コンテンツの動向に目を向ければ，PC の低価格化，HDD の低価格化，大容量化，ブロードバンドの普及，そしてホームビデオの普及によって，個人の映像の楽しみ方の可能性が広がりつつある．こうした中，個人が様々な種類の非常に大量のコンテンツの中から，どのようにしてそのユーザが映像情報の管理を行うか，自分の好みの映像を探すか，または効率的に映像を概観するかといった，映像の管理，検索，閲覧等様々な問題が生じつつある．個人で取り扱う映像の量が増えたとしてもその映像を扱う環境が整っていなければ，膨大な映像資産を十分に生かす事ができないのである．

そこで本研究では，近年急速に普及しつつあるホームビデオ映像を対象に，上記の問題点の解決を目的とした映像インデクシング手法を提案する．さらに，ホームビデオ映像の特徴を明らかにすることで本研究の目的を明確にする．ホームビデオ映像には放送映像とは本質的に異なる点がいくつか存在するため，放送映像を対象にした映像インデクシング手法がそのまま適用できない場合も多い．自動的なインデクシング手法の確立は貴重な個人の記録であるホームビデオ映像のより効果的な取り扱いを可能にするであろう．

ホームビデオの放送映像との最も重要な相違点は，ホームビデオ映像が個人で取り扱う限り，商用を目的としていない点である．これはつまり，映像の編集や検索のためのメタデータの付与をユーザ本人の責任で行わなければならないことを意味している．放送映像では，人を雇い，十分な資金を用いてこれらの作業を効率的に集中的に行うことができる．しかし，個人でホームビデオ映像のためにそこまでの努力や，資金，時間を消費することのできる人は稀であろう．ホームビデオで撮影されるイベントは，たとえば日常生活の様子，結婚式や，旅行，または会議の様子，プレゼンテーションの記録等が考えられる．会議の様子や，プレゼンテーションの記録は個人で楽しむ思い出の記録ではないが，商用を目的としなければ編集や検索，閲覧，管理のために費やすことのできる時間や，資金，人手は限られたものになるであろう．

ここで，ホームビデオの放送映像に対する相違点は次の３つの点にまとめられる．

１）ホームビデオ映像は一般に冗長である
２）放送映像と異なり決まった構造を持たない

３）映像の情報管理を個人で行わなければならない

１）に関しては，未編集の映像は映像の流れが洗練されておらず，全体の内容がつかみにくいといえる．効果的に映像のストーリーをまとめて，わかりやすく編集するには多大な労力を必要とする．また，２）に関してはたとえば，スポーツ映像やニュース映像がわかりやすいだろう．これらの映像は開始から終了まである一定の規則に則った映像となっている．野球であれば，ピッチャーの後ろ側から映した映像，バッターの映像，ヒット，ホームランなどイベントを定義することで，その映像の内容を統一的にまとめることが可能である．ニュース映像においても，アンカーショットから記事の内容の映像に移り，またアンカーショットに戻るといった映像の流れが定義できる．しかし，ホームビデオ映像という一般的な映像を取り扱う場合は，こういった構造を定義できないため，ホームビデオに適したイベントの定義，インデックスの定義が必要であろう．さらに３）では，ホームビデオは個人で取り扱う映像メディアであるため，web や雑誌から映像の情報を取得できない，何らかの機関から映像に関する情報提供が期待できないという点がある．放送映像では，映像の登場人物や放映時期，全体の概要等の情報は放送局，web 等から取得することが可能である．

これまで，ホームビデオ映像を対象にしたインデクシング手法，映像要約手法はいくつか提案されているが，音声をインデクシングの対象としたものは少ない．そこで，本論文ではホームビデオ映像の中でも特に音声情報に注目したインデクシング手法について提案した．また，これまで様々な音声特徴量が提案されており，それらの特徴量における本論文で定義する音声イベントの特性を明らかにする．なお，本論文で提案する手法はホームビデオ映像の効果的な利用を目指したものではあるが，音声情報自体に放送映像，ホームビデオの差はないため放送映像に適用することも可能である．

本論文で提案する手法は，判定ルールに基づいた手法と Gaussian Mixtures Models (GMM) を用いた手法の２種類である．これらの手法では音声イベントとして，発話，無音，音楽，背景音の４つを定義し，インデクシングを行う．さらに，従来では一つのセグメントに対して一つのイベントを対応させる手法が主流であった．そこで，本論文で提案する手法では，これらの音声イベントを一つのセグメントに対して，複数のイベントを対応させるという意味で，層状インデクシングとして提案する．

判定ルールに基づいた手法では，特徴量と音声イベントの関係から各イベントを識別するためのルールを導き出しさらに，１秒間のセグメントに対して複数の音声イベントを対応づける層状インデクシングを実現する．概要としては，発話，音楽，背景音それぞれのトレーニングデータを用いて特徴量を求める．これらの特徴量は音声イベント(発話，音楽，背景音)毎に異なった分布を持っているため，あるルールを適用することにより音声イベントの判定が可能となる．本手法では，これらの特徴量の音声イベントとの関連に基づいて，発話，音楽，背景音を検出するためのルールをそれぞれ導いた．このルールの出力は，音声イベント毎に独立して得られるので層状

インデクシングが可能である．なお，本手法では無音の検出に関しては STE (Short time energy) の閾値処理を用いている．最終的な層状インデックスはそれぞれの音声イベント判定ルールの出力に対して閾値処理を行い，閾値よりも高い値を持つセグメントに対してインデックスを付与する．

実験ではテストデータとして混合のないイベント，複数の音声イベントが混合しているデータに対して本手法を適用し，本手法の性能を検証した．また，実際のホームビデオで録音された音声を用いて層状音声インデクシングを行った．実験ではテストデータとして混合のないイベント，複数の音声イベントが混合しているデータに対して本手法を適用し，本手法の性能を検証した．また，実際のホームビデオで録音された音声を用いて層状音声インデクシングを行った．

実験により，各音声イベントが重なりをもたないテストデータでは本手法により適切な検出を行うことができることを確認した．しかし，実際のホームビデオデータでは発話に関しては良好な検出がされたが．しかし，音楽，背景音に関しては適切な検出もされているが誤検出が多くあった．

次に，ＧＭＭを用いた層状音声インデクシングを提案する．先に判定ルールに基づいた手法を提案したが，本手法では発話，音楽，背景音をトレーニングデータとし，ＧＭＭに学習させることで統計情報に基づいた手法について検討する．判定ルールに基づいた手法は発見的な手法であった．そこで，本手法では GMM を用いてトレーニングデータの特徴量の統計的な情報に基づいた手法を提案する．概要としては，まず発話，音楽，背景音のトレーニングデータに対して，GMM パラメータの学習を行う．GMM のパラメータの学習には EM アルゴリズムを用いた．次に，学習された GMM に対して，テストデータを適用し尤度を求める．最終的な層状インデックスは尤度に閾値処理を適用することで確定を行う．実験データには重なりをもたないテストデータ．重なりを持たせたテストデータ，実際のホームビデオで録音された音声を用いて本手法の性能を検証した．

本手法では重なりを持たないデータに対しては，ほぼ良好な結果を得ることができた．しかし，ホームビデオデータに対しては判定ルールに基づく手法と同様に音楽，背景音の適切な検出もされたが誤検出も多くあった．

また，本論文ではホームビデオ映像に対する映像ブラウザの試作も行った．本ブラウザでは音声イベントの発生毎にキーフレームをサムネイルとして表示し，映像中の任意のショットにアクセス可能である．任意のショットの再生中にはどのイベントがそのショット内に存在するかを表示することにより，ユーザにインデックスと映像内容との対応関係を提示している．

以上のように本論文では，放送映像とホームビデオ映像の比較を行うことにより，ホームビデオ映像の特性を明らかにし，従来提案されてきた音声特徴量における本論文で定義した音声イベントの特性を示した．また，音声に基づいた層状インデクシング手法を，判定ルールに基づく手法と GMM に基づく手法につい提案し，映像ブラウザの試作を行った．