映像シーンを簡単に検索できる“顔deナビ™”

Easy Video Scene Search using "Face Navigation"

久保田　英俊　■ KUBOTA Hidetoshi
桃崎　浩平　■ MOMOSAKI Kohei
青木　恒　■ AOKI Hisashi
風間　久　■ KAZAMA Hisashi

デジタル映像機器が急速に普及し，家庭においても大量の映像が蓄積されるようになったことで，視聴したい場面を効率よく探し出すニーズが高まっている。東芝のAVノートPC（パソコン）Qosmio™では，メディア ストリーミング処理プロセッサSpursEngine™の優れた画像処理能力を活用して，映像のシーンを簡単に検索できる“顔deナビ™”機能を搭載した。

顔deナビ™では，当社の映像インデクシング（映像内容解析）技術を適用することにより，ビデオの見どころを一覧表示したり，出演者の顔や音響シーンの変化をグラフィカルに表示する機能を提供している。

Demand has been increasing for effective searches of large volumes of accumulated data recordings due to the progress of digital video devices. In response to this trend, Toshiba has developed the "Face Navigation" function to make it easier to find scenes in a video utilizing the advanced video processing performance of the SpursEngine™ stream processor in the new Qosmio™ audiovisual (AV) notebook PCs.

"Face Navigation", based on our advanced video indexing technology, makes it possible to display a graphical view of highlight scenes, faces and audio scene changes in a video content.

1　まえがき

家庭内の映像情報のデジタル化が進むなか，AVノートPCを含めたデジタルAV機器では，長時間で大量の映像データの中から，ユーザーが見たい映像コンテンツを簡単に探せるようにすることが課題の一つになっている。

東芝は，映像認識技術を駆使した今までにないインデクシング機能を活用して，映像のシーンを簡単に検索できる“顔deナビ™”機能を開発した。顔deナビ™では，出演者の顔を一覧表示する“顔サムネイル表示”や，番組のシーンを一定間隔ごとに小さな画像にして一覧表示する“じゃばらサムネイル表示”，映像の盛り上がりのシーンを歓声や音量のレベルで表示する“音量レベル表示”，場面ごとの音の雰囲気を色分けする“音響シーン表示”などにより，映像ファイルの概要が一目でわかるようになっている。また，出演者の顔をクリックするだけでシーンを再生することもでき，見たいシーンを簡単に選ぶことができる。ここでは，顔deナビ™の機能とともに，映像インデクシング技術の詳細について述べる。

2　顔deナビ™の機能

顔deナビ™の画面の一例を**図1**に示す。

顔サムネイル表示は，等間隔で分割した時間ごとに抽出された顔画像（正方形）を並べて表示している。サムネイルのサイズと数は3種類から選択でき，整列方法は，時系列に関連シーンの人物が集まる“時間”，主要人物が見やすい“頻度”，及び人物以外の情報が役だつ“シーン混在”の三つを用意している。なお，人物の顔が少ない場合は，画面全体のシーンサムネイル（横長）で補完される。

図1. 顔deナビ™画面例 — 顔サムネイル，音量レベル，区間バー，じゃばらサムネイルで構成される。
Example of index viewer of "Face Navigation"

音量レベル表示は，比較的単純な音量のグラフ（だいだい色）の手前に，歓声や拍手のグラフ（黄色）を重ねて表示している。区間バーは，“音楽／CM”と“音響シーン”の2種類の色分けのモードを切り替えて表示する。

画面下部のじゃばらサムネイル表示では，映像から等間隔で抽出した静止画を，注目箇所の画像が開いた蛇腹状に配置し，マウスやリモコンの操作で高速に映像内を閲覧できる。

3　映像インデクシング ソフトウェアの構成

Qosmio$_{TM}$に搭載されている映像インデクシング処理に関係するアプリケーションソフトウェアの構成を**図2**に示す。国内モデルのテレビ（TV）用統合AVアプリケーション Qosmio AV Centerでは，設定によって“地デジ8倍録画”などのトランスコード機能を使った録画と同時に，インデクシング処理が実行できる。これらはSpursEngine$_{TM}$で高速に実行されるので，録画終了後に長時間待たされずに顔deナビ$_{TM}$機能が利用できる。一方，Ulead DVD MovieWriter®(注1) for TOSHIBAを利用して取り込んだビデオカメラの映像や，海外モデルにおいてWindows®(注2) Media Centerで録画されたTV番組などは，“東芝グラフィカル ビデオライブラリ”が担当する。

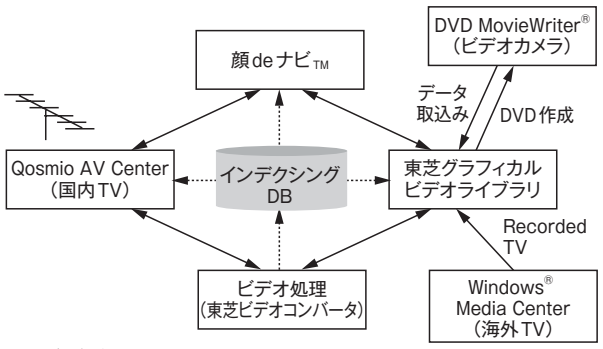

図2．アプリケーションの構成 — 顔deナビ$_{TM}$とビデオ処理が複数のアプリケーションから利用される。

Configuration of applications related to video indexing

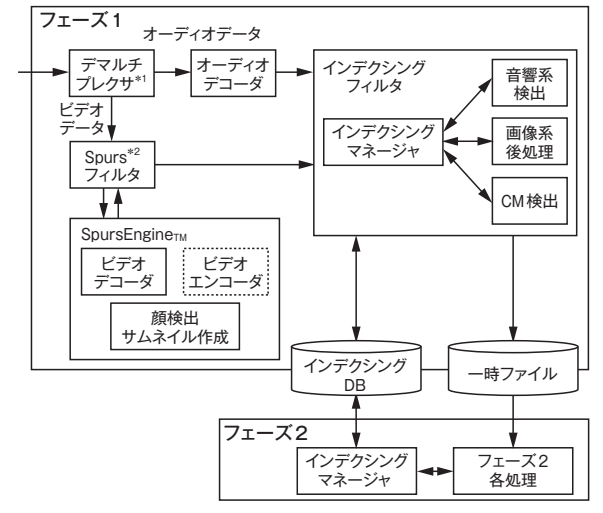

図3．インデクシング処理の構成 — SpursEngine$_{TM}$を使って映像ストリームを処理するフェーズ1と，一時ファイルを処理するフェーズ2から構成される。

Flow of video indexing process

インデクシング処理結果はコンテンツと関連付けてデータベースに格納されており，顔deナビ$_{TM}$のほか，Qosmio AV Centerや東芝グラフィカル ビデオライブラリの再生画面からも利用でき，コマーシャル（CM）と本編の境界の情報を利用したスキップ再生機能などを実現している。

インデクシング処理は，番組録画などの映像の入力に同期して処理するフェーズ1と，終端に達してから処理するフェーズ2から構成され（**図3**），地上デジタル放送のような著作権が保護された映像を扱うため，Windows Vista$^{TM(注3)}$の新しいマルチメディアプラットフォームであるMedia Foundationの保護環境で動作する。また，画像を処理するSpursEngine$_{TM}$部分と，主に音響を処理するCPU部分とが連携して動作する。

4　顔サムネイルによるシーン内容の一覧表示

4.1　処理の流れとSpursEngine$_{TM}$の利用箇所

映像から出演者の顔を検出するインデクシング処理は，フェーズ1として録画時にトランスコード処理と同時に行われる。フェーズ1の処理は，音響系の処理と映像系の処理に分かれる。

映像系の処理は，映像の復号及び符号化（トランスコード処理），顔検出処理，及び顔サムネイルとシーンサムネイルの作成で，処理性能と速度の両立が求められるため，SpursEngine$_{TM}$を利用する。SpursEngine$_{TM}$を利用して映像系の処理を担当するライブラリを顔検出ライブラリと呼ぶ。

4.2　顔検出ライブラリに求められる機能及び性能

フェーズ1のインデクシング処理は，顔検出を含めて，リアルタイム処理が必須である。顔deナビ$_{TM}$を実現するためには，リアルタイム処理を実現しつつ，時間的，空間的に，緻密（ちみつ)なデータを生成できる速度性能が必要である。なぜなら，登場人物がカメラの方を向くのは一瞬かもしれないし，抽出した顔情報からカットの解析を行うためには，顔の向き，大きさ，及び位置が滑らかに変化する検出結果が期待されるからである。また，登場人物の重要度は検出の時点では判断できないので，画像内に多数の顔が存在すれば，すべてを同時に検出しなければならない。

更に，映像インデクシングのための顔検出では，デジタルカメラや入退出管理システムで求められる顔検出と異なり，機能的には次のことが求められ，より広い条件で検出性能を維持しなければならない。

（1）　顔の向き，大きさ，位置などを被写体側が調整することがないので，より多様な顔の像を検出すること。

（2）　演出などのために，照明条件や背景の変化パターンが

（注1）　Ulead及びDVD MovieWriterは，Corel Corporation又はその関連会社の商標又は登録商標。

（注2），（注3）　Windows及びWindows Vistaは，米国Microsoft Corporationの米国及びその他の国における商標又は登録商標。

多いため，より多様な条件で検出能力を安定化させること。

顔検出ライブラリには，このほかに，シーンサムネイル生成と顔サムネイル生成の機能が求められる。

以上を踏まえ，顔検出ライブラリの目標性能は，100 ms以下の間隔の画像から同時に15個までの顔を検出し，そのほかの処理を含めて，リアルタイム処理を実現することとした。

4.3 SpursEngine$_{TM}$内の実装と最適化

SpursEngine$_{TM}$内には4個のSPE（Synergistic Processor Element）が搭載されており，担当する処理を並列化実装する。並列化実装では，効率向上のために4個のSPEに均等にタスクを配分することと，処理ステップをパイプライン処理化して並列実行できるようにすることが重要である。今回の顔検出ライブラリでは，コアごとに異なる処理を分担する並列化モデルと，マスタワーカ型と呼ばれる並列化モデルを組み合わせた方式を採用した。4個のSPEの分担は以下のようになる。

(1) SPE0（コーデック制御SPE）　デコーダハードウェア（HW）及びエンコーダHWの制御と，画像処理のために画像を正規化する処理などを担当する。

(2) SPE1（顔検出マスタSPE）　顔検出マスタとして顔検出の実処理を小タスクに細分化してワーカに分配する機能と，顔検出結果から顔サムネイルを生成する処理，シーンサムネイルの生成などを担当する。

(3) SPE2，SPE3（顔検出ワーカSPE）　顔検出マスタSPEよって細分化された小さなタスク（顔検出小タスク）を，タスクキューから取得して処理することを繰り返す。顔検出の主たる処理はワーカSPEが担当する。

SPEごとに異なる処理を振り分けつつ，4個のSPEとしてはパイプライン処理が実現されるように実装している。

4個のSPEの分担とデータの流れを**図4**に示す。

この並列化方式では，SPE0，SPE1の処理量は入出力の映像ストリームのデータ量に依存するものの，映像の内容への依存性は高くない。一方，SPE2，SPE3の処理量は映像の内容への依存性が高く，画像内に映っている顔の数に応じた処理量の時間的変化がある。処理量の時間的変化はパイプライン処理が滞る原因になるが，マスタワーカ型の実装であれば，タスクキューがバッファとして働くため，全体としての処理効率は落ちない。この設計モデルの採用により，悪条件の画像でも，全体としてリアルタイム処理を確保することができた。

SPE2，SPE3に実装する顔検出処理は，ハンドジェスチャ認識処理とアルゴリズムを共通化して，最適化と並列化を集中して行った。その結果，顔検出ライブラリは，機能要求を実現したうえで，トランスコード処理と同時にインデクシング処理をリアルタイムで実現することができた。

ハンドジェスチャ認証のアルゴリズムとその最適化については，この特集の論文“ハンドジェスチャ インタフェース技術”(p.58－62参照)で述べる。

4.4 顔サムネイルの作成

顔検出ライブラリによって出力されるのは1フレームごとの顔の位置情報であるため，顔追跡を行い，連続する複数フレーム間で位置とサイズの変動が小さいものを同一被写体の顔と推定する。しかし，TV放送など編集済みコンテンツでは，①カメラ切替わり点の前後の連続するフレームで，別の被写体の顔の位置やサイズがほぼ同じである場合，②別のカメラから再び元のカメラに切り替わり，時間的に不連続のフレームにも同一被写体の顔がある場合，に対処する必要がある。

そこで，当社のハードディスク＆DVDレコーダ搭載の“マジックチャプター”機能を実現しているカット検出と類似ショット検出の技術を用い，複数フレーム間での顔領域が同一であるか否かを高速に判定する。

まずフレームごとに，画面全体の色調と輝度配置パターンから成る特徴量を計算し，その特徴量が急激に変動する点をショット切替わり（カット）と推定する。また，時間的に離れたフレーム間で同様の特徴量を持つ場合には，それらのショットは類似である，すなわち，同じカメラから撮影された可能性が高いと推定する（**図5**）。

全体としては，①顔追跡により，連続したフレーム間で顔領

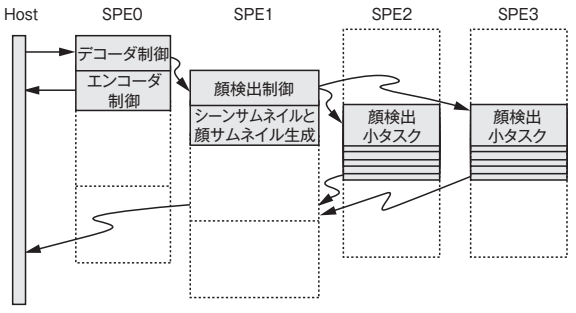

図4. 四つのSPEによる並列化 — 四つのコアで異なる処理を担当し，かつパイプライン処理を実現する。
Parallel and pipeline processing by four synergistic processor elements (SPEs)

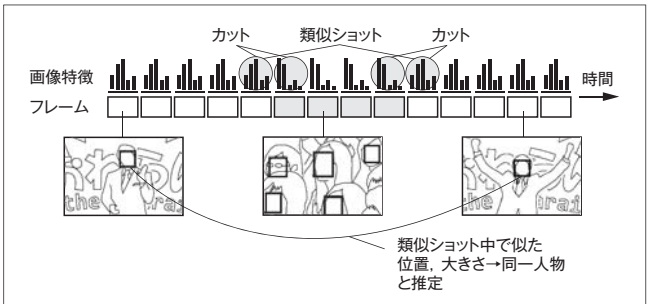

図5. 同一被写体の推定 — 顔検出と類似ショット検出の組合せによって，時間的に離れた位置にある顔領域を同一人物と推定する。
Recognition of same person's face from size and position in each frame

域の位置とサイズの変動が少ないものをグルーピングし，②カット点が到来したらグルーピングを打ち切り，③類似ショット間で同様の位置とサイズの顔領域は再びグルーピングする，という処理により，顔領域が同一被写体のものであるかどうかを推定する。推定された同一被写体グループそれぞれから顔サムネイルを生成することにより，計算処理量を抑えつつ，同一人物の顔サムネイルが大量に出力されることを低減している。

5　音響解析によるシーン内容の一覧表示

5.1　音響情報に表出するシーンイベント

様々なテレビ番組やホームビデオ映像に共通するイベントとして，歓声，拍手，音楽が挙げられる。また，演奏ボリュームの変動や笑い声による音量レベルの増減をそのまま画面に示すことによっても，ユーザーは映像中で盛り上がっている場面を知る手がかりとなる。これら歓声，拍手，音楽，音量レベルは上述の顔サムネイルやシーンサムネイルでは示せないため，これらを示す時間軸の区間バーと棒グラフの表示を導入した。

5.2　歓声，拍手，音量レベル，音楽区間の表示

棒グラフ表示のうち，音量レベルは音響信号波形の二乗和を用いて単純計算し，盛り上がりバーとして表示する。歓声及び拍手については，音響信号にスペクトル分析を行い，あらかじめ登録されたスペクトルモデルと比較することで“歓声らしさ”や“拍手らしさ”を算出する。これらを重み付け平均した結果を歓声バーとして表示する。

音楽区間は音楽番組だけでなく，ドラマのオープニング，学芸会や結婚式の出し物部分など，幅広いジャンルで意味のあるイベントである。そこで，音楽演奏に特有の周波数パターンを探索するとともに，大量のデータから事前に学習させた音楽及び音楽以外の音響モデルとの重み付け比較を行って真の音楽らしさを算出し，算出結果が一定値以上である部分を音楽区間と推定する。

5.3　音響シーンの表示

情報番組などでスタジオトークと情報ビデオが交互に編集されているような場合においては，トーク，ビデオなどの番組構成の変化を表示することで内容の理解を支援できる。そこで，場面ごとの音の雰囲気を色分けして表示する音響シーン クラスタリング技術を導入した。

この技術では，音声認識でも用いられる話者交代検出技術を応用し，発話者が交代するたびにシーンの切替わりを定義する代わりに，複数の話者から構成される大局的な音響特徴量を算出し，その特徴量の変動部分で区間バー上の色を変化させるとともに，類似した特徴量を持つ区間を同色で表示する。これによって，例えばスタジオトークと情報ビデオが切り替わっている時点や，情報ビデオ後に再びトークに戻っていることなどを，色の違いでユーザーが判断することができる（**図6**）。

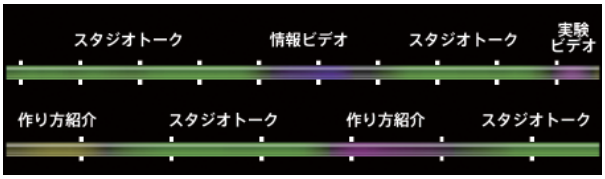

図6. 音響シーンクラスタリング — 情報番組（上段）と料理番組（下段）の解析結果で，目盛りは1分を示す。場面の切替わりが色分けされ，同一の意味を持つ場面は同色で表示される。

Results of audio scene clustering

6　あとがき

顔deナビ$_{TM}$機能の概要と，利用される映像インデクシング技術について述べた。この機能によって，映像コンテンツの内容を一覧し，シーンを選んで見るという新しい視聴方法を実現できた。また，SpursEngine$_{TM}$を使った，トランスコードとインデクシングを同時に行うアプリケーションを実現できた。

今後は，映像コンテンツ間を横断した検索機能や，今回の顔検出から顔認識を応用した機能拡張を行うことで，映像インデクシング機能の強化を目指していく。

久保田　英俊　KUBOTA Hidetoshi
PC＆ネットワーク社 PC開発センター PCソフトウェア設計第二部グループ長。AV PCソフトウェアの開発に従事。
PC Development Center

桃崎　浩平　MOMOSAKI Kohei
PC＆ネットワーク社 PC開発センター PCソフトウェア設計第二部主務。AV PCソフトウェアの開発に従事。情報処理学会，日本音響学会会員。
PC Development Center

青木　　恒　AOKI Hisashi, Ph.D.
研究開発センター マルチメディアラボラトリー主任研究員，工博。映像インデクシング，ヒューマン・コンピュータ・インタラクションの研究に従事。情報処理学会会員。
Multimedia Lab.

風間　　久　KAZAMA Hisashi
セミコンダクター社 システムLSI事業部 先端SoC開発センター参事。画像処理技術，画像認識技術の開発に従事。電子情報通信学会会員。
System LSI Div.