オーディオビジュアル複合情報処理 30－5
（2000．9．22）

動きに基づくいくつかの映像処理手法

吉田 俊之

東京工業大学大学院理工学研究科集積システム専攻
〒 152-8552 目黒区大岡山 2-12-1
E-mail : tyoshida@ss.titech.ac.jp

あらまし　ハードウエア技術の急速な進展に伴い，今日ではパーソナルコンピュータ上であっても動画像の記録，再生等が可能となり，またインターネットの普及ともあいまって，誰もが動画像を広く公開することができるに至っている．このため，動画像に対する管理，編集，加工等の技術は，従来にも増して重要性を帯びてきている．本稿では，特に動画像から得られる動き情報に着目し，これを利用した動画像検索手法や，動画像中の対象物体に対するセグメンテーションおよびトラッキング手法に関する著者らの研究について触れさせて頂く．

A Review on Video Processing Techniques based on Motion Information

Toshiyuki YOSHIDA

Dept. Communications and Integrated systems
Graduate school of Science and Technology
Tokyo Institute of Technology

2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan
E-mail : tyoshida@ss.titech.ac.jp

Abstract　Rapid progress of hardware technology and popularization of the Internet has enabled us to store motion images even on a personal computer and open them to the public. Techniques on video management, editing, manipulation, and so on are thus becoming more and more important. This paper reviews the authors' research on video processing techniques based on motion information ; video retrieval, target object segmentation and tracking.

1. はじめに

今日では，ハードウエア技術の進展に支えられた計算機処理能力の向上，記録装置の大容量化等によって，従来では巨大なハードウエアを持ってしてのみ可能であった動画像の記録，編集，加工等が，個人ベースのパーソナルコンピュータ (PC) 上であっても可能となるに至った．また，従来のテープメディアに代わって高速ランダムアクセス可能なハードディスク等への動画像の記録が可能となり，動画像へのアクセシビリティも格段に向上している．さらに，インターネットの普及とその高速大容量化により，誰もがプライベートな映像を公開することが可能となっている．

以上の背景の下，動画像は我々にとって益々見近なものとなり，特にコンピュータ上で動画像を扱う機会が増えている．従って，溢れる動画像情報を管理，検索したり，逆にプライベートな映像を加工・編集して公開するなどの作業を支援する映像処理技術に対する要求は益々高度化するものと考えられる．従来は研究室レベルで研究されていた種々の映像技術が，我々のPC上のソフトウエアとして実用化される下地は整いつつあると考えられる．

動画像とは，結局のところ静止画を時間的に並べたものであるが，単なる静止画集合と決定的に異なるのは，動画像の持つ「動き情報」である．これを積極的に利用し，上述のような動画像処理アルゴリズムを開発することは，一つの重要な研究分野である．本稿では，以上のような「動き情報」を利用した映像処理手法のうち，動画像検索手法，および動画像中の対象物体のセグメンテーション・トラッキング手法に関する著者らのアプローチについて述べる．

以下，まず 2. では，動きに基づく MPEG 動画像の検索手法について述べる．MPEG は動画像圧縮符号化アルゴリズムとして最も広く利用されている手法で，フレーム間相関の除去に動きベクトル (MV) ／動き補償を用いているため，この MV を動画検索に応用しようとする試みは非常に素直な考え方である．しかしながら，これには MV の精度という解決すべき問題があり，通常の MPEG エンコーダによって推定される MV を用いたのでは高い検索精度は期待できない．そこで，著者らは，まず MPEG の符号化効率を大きく損なうことなく極力対象物体の動きに忠実な MV を推定する手法を開発し，この手法を用いて符号化した MPEG 画像に対して階層的な検索手段を提供する動画像検索手法を研究している．これらの手法についてその概要を説明する．続く 3. では，動画像中の対象物体のセグメンテーション・トラッキングに関する一手法を紹介する．動画中の対象物体を切り出す手法としては，まず基準フレーム内の対象物体を高精度に切り出し，これを時間的にトラッキングしていくアプローチが主流である．そこで，まず新しいエッジ画像の補間強調手法とそれを用いた静止画像セグメンテーション手法を紹介した後，Watershed 法を 3 次元に拡張して物体トラッキングを行なう手法について触れる．最後に，4. で小文を総括する．

2. 動きに基づく MPEG 動画像の検索

1. で述べた通り，動画像は「静止画像の集合」と「動き情報」の和と解釈できるため，動画像を検索する際には，一般に，

1. シーンチェンジ検出手法を利用し，動画像を複数のシーンに分割する．
2. 各シーン内で代表的なフレーム（一般にキーフレーム等と呼ばれる）を抽出する．
3. シーン内の動き情報を抽出する．
4. キーフレームに対しては，静止画像の検索手法を適用し，同時に動き情報に基づく検索を併用し，目的とするシーンを検索する．

というアプローチが取られている [1]．本章では，このうちの特に動き情報に基づく検索手法についての著者らのアプローチを紹介させて頂く．

2.1 MV の高精度推定手法

周知のように MPEG は動き補償と DCT を併用した符号化方式で，その MV を相関の除去以外に応用しようとする試みは数多く見られる．しかしながら，MV は本来的に画像の動きを表すベクトルではなく，むしろフレーム間相関を最も効果的に除去するベクトルであるため，相関除去以外の目的に応用する際には，MV の精度が問題となる．実際，MPEG に関する種々の疑問点をまとめた MPEG FAQ[2] において，次のような記述が見られる．

- Can motion vectors be used to measure object velocity ?

  Motion vector information cannot be reliably used as a means of determining object velocity unless the encoder model specifically set out to do so. First, encoder models that optimize picture quality form vectors that typically minimize prediction error and, consequentially, the vectors often do not represent true object translation. 中略. Secondly, motion vectors are not transmitted for all macroblocks anyway.

そこで，著者らはまず，ここで言うところの "encoder model specifically set out to do so" を実現するため，符号化効率を大きく損なうことなく，極力実際の対象物体の動きに忠実な MV を推定する手法の構築を試みた [3]．本節では，この手法の概要について述べる．

MPEG エンコーダにおける動き推定には，ほぼ例外なくブロックマッチング (BM) 法が用いられる．図 1 は単純 BM 法によって推定した MV の例

で，図2は，図1の(a)〜(e)の各ブロックに対して，著者らの提案するMVの信頼度関数[4]

$$R(\theta) = C \cdot \frac{\sin^2\theta \int {f_x}^2 dx + \cos^2\theta \int {f_y}^2 dx - 2\sin\theta\cos\theta \int f_x f_y dx}{\int {f_x}^2 dx \int {f_y}^2 dx - \left(\int f_x f_y dx\right)^2} \quad (1)$$

を計算した結果である．式(1)は，θ方向の推定誤差の2乗平均を与える式で，図1，2と共に観察すると次のことが解る．

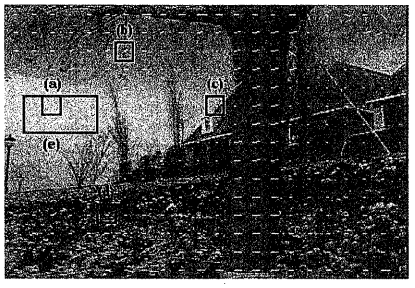


図1：単純BMによるMVの推定例

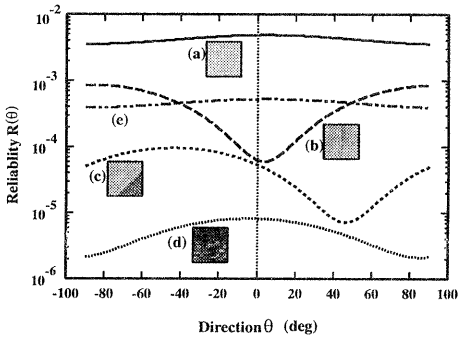


図2：図1の(a)-(e)に対する信頼度関数$R(\theta)$

- (a)のようなフラットなブロックでは$R(\theta)$はあらゆる方向に大きな値を持つ．
- (b)，(c)のようなエッジブロックでは$R(\theta)$はエッジに沿った方向に大きく，それと直交方向に小さな値を持つ．
- (d)のように特徴的なパターンを含むブロックでは$R(\theta)$はあらゆる方向に小さな値を持つ．
- フラットなブロックであっても(e)のように大きなブロックを取れば(a)の場合よりも$R(\theta)$の値は小さくなる．

$R(\theta)$はθ方向の推定誤差であるから，その値が小さいほど推定が容易となる．すなわち，(a)のように単独では推定が困難であるブロックも周囲の同種のブロックと統合し，大きなブロックで推定を行なうと精度の向上が見込まれる．また，$R(\theta)$の大きさおよびその最大値と最小値の比を用いることにより，フラットブロックとエッジブロックを分類することも可能である．

そこで，(1)通常のBMを実行した後，$R(\theta)$を用いてブロックを分類し，(2)周囲の同種ブロックを統合した大きなブロックについて(3)再度BMを実行する，というアルゴリズムにより，特に従来の単純BMでは推定の難しかったフラットな部分と単純エッジを持つ部分の推定精度の向上が図れる．一方，BMがMV推定に失敗するのはこの他に，並行移動以外の大きな動きを持つブロック，およびオクルージョン／アンカバード領域であるが，これらについては基本的に並行移動量を推定することは困難であるため，MVの推定は行なわず，MPEGエンコーダに積極的にイントラ符号化させる．

以上により，MVの推定精度は改善されるが，一方でMPEGに応用した際に符号化効率の低下が懸念される．しかしながら，いくつかの実験の結果，通常のMPEGエンコーダと，本手法を用いたエンコーダ（MVのポスト処理は行なわない）の符号化効率を比較すると，MVのコヒーレンスのため効率が向上する場合もあり，平均して本手法を用いた場合の効率低下は3〜5%程度であった．図3は，特に改善が著しい，フラットな部分を多く含む画像に対する推定結果である．

図3：単純BM法と本手法によるMVの比較

本手法を用いて，高精度，高コヒーレンスかつスキップブロックを極力減らしたMV場を与えることにより，MVを利用する種々のアプリケーションの大幅な精度向上が見込まれる．

2.2 動きに基づくMPEG動画像検索

検索システムを設計する際には，まず用いる検索キーを検討する必要がある．我々に取って最も扱い易いのはキーワードを用いた検索であるが，これを可能とするには個々の検索対象データにキーワードを割り振る必要があり，動画像に対しては実用的ではない．そこで，我々は所望のシーンに近い類似シーンもしくはGUIを通してユーザが入力した抽象的シーンをキーとして，そこから抽出される動き情報を基に動画像を検索する手法を検討してきた．

動画像検索は膨大なデータ処理を要するため，マッチング処理の高速化が不可欠である．そこで，高速かつ低精度のマッチングを行なう階層から，低速である一方で高精度マッチングを行なう階層までの複数の階層を用意し，これらによる階層的な検索手法を検討した．具体的には図4に示すように，一つのシーンを時空間データと捉えてそれを適当なサイズの時空間サブブロックに分割し，各サブブロック内でMVのヒストグラムを取ることによって動き情報を抽出する．サブブロックのサイズを階層毎に変化させることで，マッチング精度とヒストグラム間距離の計算コストの間のトレードオフを実現する．

実際の検索は，GUIを通して作成したキーシーンとデータベース中の対象シーンとの間でヒストグラム間距離を測定し，最も距離の小さいものをその階層における候補とする．これを最下層から最上層へと繰り返して実行し，対象候補を絞り込んでいく．ヒストグラム間距離には通常の2乗距離を用いる．ただし，人間の動き認識に関する特性や，ユーザがキーシーンを作成する際の曖昧さを吸収するため，各階層の各サブブロックで得られるヒストグラムを平滑化した後，2乗距離を求めている．

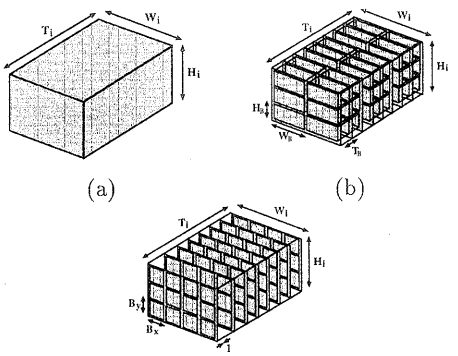


図4：階層化されたヒストグラム処理

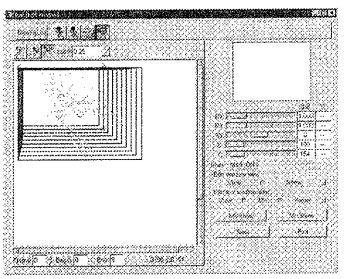
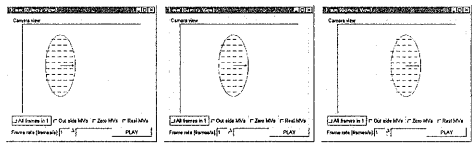
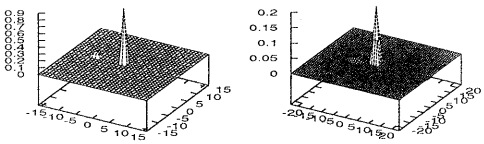


図5：キーシーン作成のためのGUIと作成例

図5(a)は，作成したキーシーンを作成するためのGUIで，ユーザは複数のフォアグラドオブジェクトの動き，バックグラウンドの動き，およびカメラワークの3つを入力することができる．一方，図6は，そのGUIを用いて作成したキーシーンの例（静止バックグラウンドに対して右方向に移動する対象物体を静止カメラで撮影）で，シーン全体を1つのサブブロックとした最下層（図4(a)）におけるヒストグラム，およびそれを平滑化した結果を同時に示す．図7は，約1200のスポーツシーンから図6のキーを用いて検索した上位3位までの検索結果で，GUIによって入力した動きに近い動きを持つシーンが得られている．

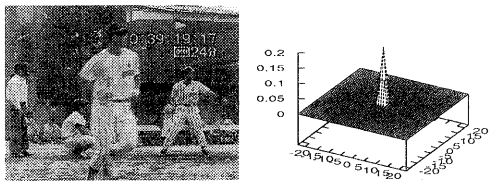


(a) 作成したキーシーン

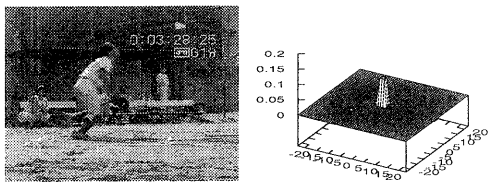


(b) ヒストグラム

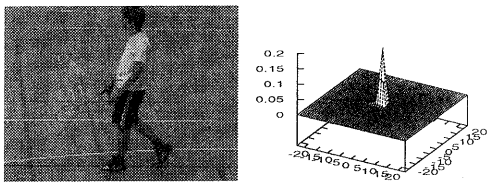
図6：キーシーンの例と最下層でのヒストグラム



(a) 第1位



(b) 第2位



(c) 第3位

図7：図6をキーとして検索した結果

3. 動画像中の対象物体のセグメンテーション・トラッキング

1. で述べたような動画像の加工・編集環境の普及や，またMPEG4に代表されるオブジェクトベースの符号化法の実用化などによって，静止画像や動画像から対象物体をそっくり抽出するセグメンテーション技術の重要性が高まっている．しかしながら，現在の計算機の処理能力や画像認識技術では，動画像セグメンテーションを完全に自動化することは困難であるため，

- ユーザがある1フレーム対象物体の概形をインタラクティブに指示する．
- 当該フレーム中の対象物体を静止画像のセグメンテーション手法によって切り出す．
- 得られた物体を時間方向にトラッキングし，対象物体を動画像として切り出す．

という手法が主に用いられている．

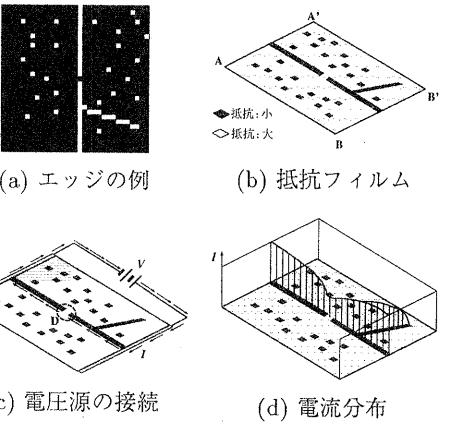
画像のセグメンテーション法としては，従来より，SNAKES，領域拡張法，メッシュを用いた手法，Watershed法など多くの手法が提案されているが，各手法固有の問題のため，実用的な抽出精度は得られていないのが現状である．本章では，セグメンテーションやトラッキングの精度改善を試みた著者らアプローチを紹介する．

3.1 電気回路シミュレーションに基づく静止画像からの対象物体抽出

静止画像の高精度セメンテーションには画素単位の処理が必要で，対象物体の完全エッジが得られれば当該処理の実現は容易となると考えられる．そこで，著者らは，何らかの物理現象に基づいてエッジを補間・強調する手法を検討し，電気回路における電流分布に着目した[8]．

図8(a)は，画像から例えばSobel等のオペレータを用いて抽出したエッジのモデルで，これを(b)に示すように，エッジ上で電気抵抗が低く，エッジ外で高くなるような分布を与えた抵抗フィルムと考えて電圧を印加すると，電流の性質から(d)に示すように，途切れたエッジ付近にも幾らかの電流が流れる一方，孤立ノイズ上にはほとんど電流が流れないという分布が得られる．そこで，この電流分布を再び輝度値に変換すると，途切れたエッジの補間と同時にノイズの除去が可能となる．

我々は，これを図9のような回路を用いて計算機上でシミュレーションを行ない，画像セグメンテーションに応用した．すなわち，エッジ画像における輝度値を単調に減少する関数を用いて図9の抵抗値にマッピングし（図中の四角形で囲った2本の抵抗が1画素に対応），電圧を印加して電流分布を求める．これを逆に輝度値に変換した後，適当な輝度値と初期点を取って領域拡散を実行し，対象物体の抽出を試みている．



(a) エッジの例　(b) 抵抗フィルム

(c) 電圧源の接続　(d) 電流分布

図8：本手法の基礎的考え方

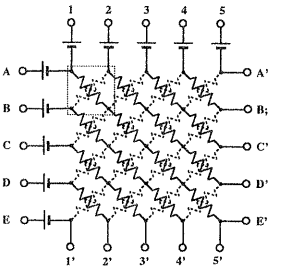
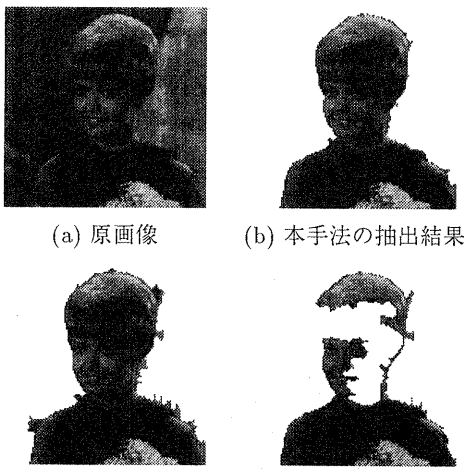


図9：シミュレーションに用いる抵抗回路網

図10は，本手法を用いて抽出した結果の例(b)，および単純な領域拡張法のみによる結果(c)，(d)である．本手法を用いることにより，エッジが補間されると共にノイズが除去されるため，(c)，(d)に比べ抽出精度が改善されている．



(a) 原画像　(b) 本手法の抽出結果

(c) 本手法のエッジ補間を用いない場合の結果

図10：本手法による抽出結果とその比較

3.2 3次元 Watershed 法による対象物体のトラッキング

次に，基準フレームに対して得られたセグメンテーション結果を用いて，これを時間的に前後にトラッキングすることにより，動画像中の対象物体を抽出する手法を検討した．具体的アイデアは，静止画像のセグメンテーション手法として広く用いられている Watershed 法を 3 次元に拡張し，時空間データを 3 次元 Watershed 法によって「背景」と「対象物体」の 2 つに分割するというものである．

まず，図 12 に示すように，基準フレームのセグメンテーション結果を物体の内部領域と外部（背景）領域に分割し，両領域の各点を図 13 のように各フレーム面内だけではなく時間方向をも含めた 3 次元 6 方向に拡散させることで，3 次元の Watershed を実現する．詳細なアルゴリズムについては文献 [9] を参照されたい．

図 14 は，“Foreman” を対象として，30 フレーム目を基準フレームに取り，前後 30 フレームの計 60 フレームに亙ってトラッキングを行なった結果である．図中では，16，30，および 44 フレームの結果を表示している．対象の人物は，左右に多少の動きを持っているものの，提案アルゴリズムによってある程度の精度でトラッキングされている様子が観て取れる．各フレームのエッジに対して前節の電気回路シミュレーションに基づくエッジ補間を施すなど，さらに精度を向上する工夫を加えることで実用に供するアルゴリズムと成り得ると考えられる．

4. おわりに

小文では，動画像中の動き情報に着目したいくつかの映像処理手法について，著者の行なってきた研究を中心に紹介させて頂いた．これらは，今後の動画像処理環境の普及と性能向上，さらにはネットワークの高速化等によって，益々実用化への要求が高まるものと考えられる．一方で，これらの分野は，常に新しいアイデアを容易に試すことのできる分野でもあり，多くの研究者の方々の斬新なアイデアを参考にさせて頂きつつ，我々も新しい手法の創造に従事していきたいと考えている．

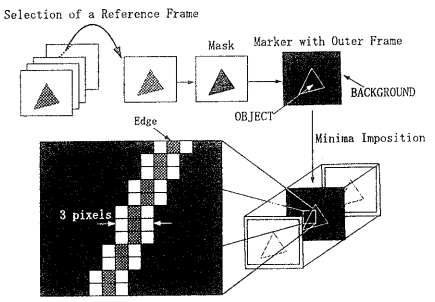


図 12：marker の作成

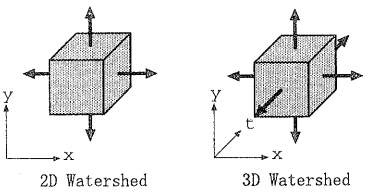
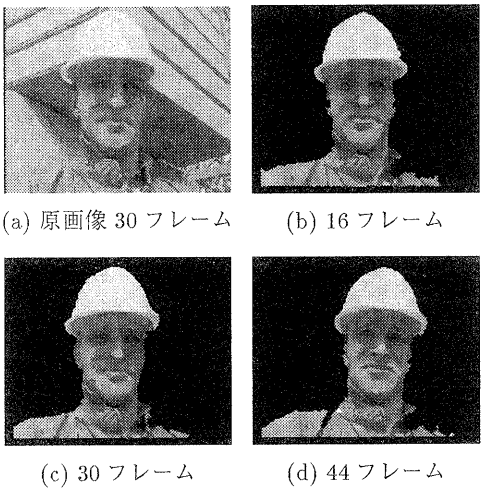


図 13：2 次元および 3 次元 Watershed 法



(a) 原画像 30 フレーム　(b) 16 フレーム

(c) 30 フレーム　(d) 44 フレーム

図 14：“Foreman” に対する適用結果

参考文献

[1] 外村ほか：“ネットワーク映像メディア”，映像情報メディア年報，映メ誌，vol. 54, no. 7, pp. 998–1004（2000 年 7 月）．

[2] http://www.crs4.it/HTML/LUIGI/MPEG/mpegfaq.html

[3] 西原，吉田，酒井：“信頼度に基づくブロック統合を用いた動画像の高精度動きベクトル推定とその応用”，映メ誌，vol. 53，no. 1，pp. 148–156（1999 年 1 月）．

[4] 吉田，宮本，酒井：“動画像の動きベクトルに対する信頼度関数とその応用”，信学論 D-II，vol. J80D-II，no. 5，pp. 1192–1201（1997 年 5 月）．

[5] 吉田：“動き情報に基づく MPEG 画像の階層的検索手法”，信学技法，IE99-80, pp. 39–46（1999 年 11 月）．

[6] Khanh V.D. and T. Yoshida：“Precise Estimation of Motion Vectors and its Application to MPEG Video Retrieval”, Proc.ICIP'99 (Oct. 1999).

[7] 大野，吉田：“動きベクトルを用いた階層的 MPEG 画像検索”，信学技法，IE2000-27，pp. 23–31（2000 年 7 月）．

[8] 平野，吉田：“電気回路シミュレーションに基づく静止画像のセグメンテーション”，信学技法，IE2000-10, pp. 9–16（2000 年 6 月）．

[9] 下里，吉田：“時空間解析を用いた動画像中の対象物体トラッキング”，信学技法，IE2000-28，pp. 31–38（2000 年 7 月）．