MPEG 符号化情報に基づく類似シーン検出方式

片岡良治† 遠藤斉†

本論文では，MPEG 符号化情報から直接的に求められる映像の特徴情報を用いて，映像に含まれる類似シーンを精度良く検出する方式について述べる．ここでいう類似シーンとは，例えば野球中継の映像に含まれる個々のホームランシーンのように，論理的に同じ意味を持つが物理的な構成が異なるシーンを指す．類似シーンを精度良く検出できれば，それらに一括して同じタグ情報を付与できるようになり，映像データベースのインデクシング作業の効率化が図れる．提案法は，特にスポーツ映像へのタグ付け処理の効率化を狙いとしており，スポーツ映像の類似シーンに共通するカメラワークの存在に着目してシーン検出を行う．また，音声認識の分野で提案された連続 DP マッチングをカメラワーク情報の照合処理に適用することで，類似シーン毎のシーン長の違いに柔軟に対応する．実際の野球中継の映像を用いて実験した結果，提案法は従来法よりも高い適合率と再現率を提供できることが明らかとなった．

Similar Scene Detection Using MPEG Encoded Video Data

RYOJI KATAOKA† and HITOSHI ENDOH†

This paper describes a similar scene detection method using feature information directly obtained from MPEG encoded video data. Scenes are regarded as similar ones when they have the same logical meaning while each of them contains different physical data. For instance, all home run scenes in a baseball program have the same logical meaning of "home run" while each of them contains different image data. Similar scene detection is effective for eliminating trouble in making an index of a video database since it makes it possible to assign the same keyword to all detected scenes at once. The proposed method detects similar scenes based on their camera work similarity. Its main application is sports scene detection since similar sports scenes are generally captured with the same camera work. To cope with the difference of scene length among similar scenes, it adopts the continuous DP matching algorithm to compare camera work features obtained from MPEG encoded video data. It is evaluated using a broadcasting baseball program. The results show that it can provide higher precision and recall rates than traditional methods.

1. はじめに

MPEG に代表される映像圧縮符号化技術の発展により，大量の映像情報をデータベースに蓄積し管理することが現実的となった．これに伴い，利用者が映像データベースから所望のシーンをその内容に基づき検索する技術の重要性が高まりつつある．内容に基づくシーン検索を実現するためには，シーンの内容を表すタグ情報（キーワード等）を映像情報に付与する必要があるが，一般に単体でも膨大な情報量を有する映像情報に人手でタグ情報を付与するのは現実的でない．このような背景から，画像理解技術や音声認識技術を応用してタグ付け処理の自動化を目指す研究が盛んに行われているが，現状の技術で完全な自動化を図るのは困難というのが実状である[1),2),3)]．タグ付け処理に人手の介在が不可欠である現状を踏まえると，内容に基づくシーン検索を実現する上では，タグ付け処理を効率的に行う手法の確立が重要と言える．

本論文では，スポーツ映像へのタグ付け処理の効率化を主たる目的とした，MPEG 符号化情報に基づく類似シーン検出方式について述べる．ここでいう類似シーンとは，論理的には同じ内容であるが物理的な構成が異なるシーンを意味する．例えば野球中継の映像におけるホームランシーンを考えると，すべてのシーンはホームランという同一の意味内容を持つが，各々のシーンは異なる画像から構成されるので，これらは類似シーンと捉えられる．このような類似シーンを精度良く検出する方式を確立し，例えば典型的なホームランシーンを手掛かりに映像に含まれるすべてのホームランシーンが検出できるようになれば，それらに一

† NTT サイバースペース研究所
NTT Cyber Space Laboratories

括して“ホームラン”というキーワードをタグ情報として付与できることになり，効率良いタグ付け処理が実現できる．スポーツ映像はニュースやドラマなどに比べ同じ意味合いの類似したシーンを数多く含むので，その効果は特に大きい．

これまでにも映像に含まれる同一内容のシーンを検出する手法[4),5)]が提案されてはいるが，それらはどれも物理的に同一の構成をとるシーンの検出を目的としており，本論文でいう類似シーンを精度良く検出する目的には向かない．また，既存の手法では映像から抽出した色情報や音響情報の照合を通してシーン検出を実現しているが，これらの特徴情報は必ずしも類似シーンの検出のために有効とはならない．本論文では，スポーツ映像特有のカメラワークの存在に着目し，MPEG 符号化情報から求まるカメラワーク情報に対して連続 DP マッチング[6)]を適用することで類似シーンを精度良く検出可能であることを示す．

以降，2.では基本事項として，本研究で扱う映像情報とシーン検出処理の基本フローについて述べると共に，従来法の問題点とそれに対処するための本研究のアイデアを概説する．3.では，本研究で提案する MPEG 符号化情報に基づく類似シーン検出法について述べる．4.では，実際のスポーツ映像を利用した実験を通して，提案法の有効性を評価する．5.では，本論文についてまとめる．

2. 基本事項

2.1 対象とする映像情報

本研究では，映像が MPEG 符号化されてデータベースに蓄積されている状況を想定し，MPEG 符号化情報から直接抽出できる映像の特徴情報をシーン検出に有効に利用するアプローチをとる．MPEG 符号化情報は，離散コサイン変換や動きベクトルの計算など比較的高度な画像処理を通して生成されているので，これを直接解析することで色情報や動き情報などの特徴情報を，改めて画像処理を行うことなく低コストで抽出可能である[7),8),9)]．尚，MPEG 符号化形式としては，現状一般的に利用可能である MPEG-1 および MPEG-2 を対象とする．

2.2 シーン検出の基本フロー

本研究において前提とする類似シーン検出処理の基本フローを図 1 に示す．探索の対象とする映像（ターゲット映像）および検出したいシーン（サンプルシーン）を入力すると，ターゲット映像の部分区間とサンプルシーンとの類似度の関係を表す情報が出力される．

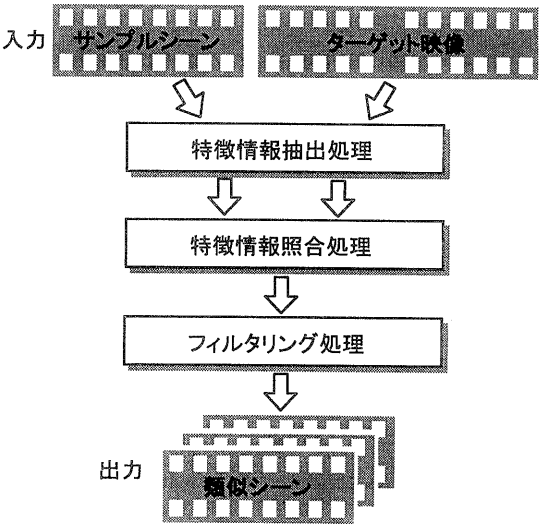


図 1 シーン検出処理フロー
Fig. 1 Processing flow of scene detection

類似度は，ターゲット映像およびサンプルシーンから抽出した特徴情報の照合処理により算出される．出力された情報を類似度に基づく閾値処理などでフィルタリングし，サンプルシーンに類似したシーンを選択する．

ターゲット映像から部分区間を選択する手法としては，画像処理などで自動的に検出できるカット点を基準に部分区間を選択する方法[10),11)]と，任意のフレーム画像を起点に部分区間を選択していく方法[4),5)]の 2 通りが考えられるが，本研究ではスポーツ映像の特性を考慮して後者のアプローチをとることとする．スポーツ映像では，有意なシーンの開始点や終了点が必ずしもカット点に対応するとは限らない．例えば野球のホームランシーンを，ピッチャーがボールを投げる場面からバッターがホームベースを踏む場面までと捉えた場合，この 2 つの場面がカメラの切り替えなどを表すカット点に対応するのは希である．

2.3 関連研究

これまでにも，映像に含まれる同一内容のシーンを図 1 の処理フローに基づき検出する方式がいくつか提案されているが，それらは照合に利用する特徴情報の種類とその照合手法の違いにより特徴付けられる．柏野らの時系列アクティブ探索法[4)]では，映像に含まれる音響信号のスペクトル特徴を特徴情報として利用する．サンプルシーンおよびそれと同じ長さで切り出したターゲット映像の部分区間についてスペクトル特徴のヒストグラムを作成し，ヒストグラムの比較により

両者を照合する．これにより，波形照合に基づく従来法に比べてシーン検出に必要な計算時間を大幅に短縮している．また，長坂らの実時間シーン分類手法[5]では，フレーム画像の色平均を特徴情報として利用する．特徴情報の時系列データを，時間軸方向の冗長性を圧縮して管理しながら，サンプルシーンと同じ特徴情報の並びを持つターゲット映像の部分区間を確実に検出する方式を提案している．

これら従来方式に共通するのは，ターゲット映像から検出すべきシーンの長さがサンプルシーンの長さと同一であることを前提としている点である．つまり，既存の方式は，映像に含まれる同じ CM を検出するというような，物理的に同一の構成をとるシーンの検出を目的とするものであり，本論文でいう類似シーンを検出する用途は考慮していない．例えばホームランの実況シーンをサンプルシーンとしてそのスローリプレーをターゲット映像から検出するというように，長さの異なる類似シーンを確実に検出できるようにするためには，特徴情報を単純に時系列順に比較する照合方式ではなく，時間軸方向の伸縮を考慮できる照合方式が必要と考えられる．

また，音響信号や色平均といった特徴情報は，物理的に同一構成のシーンを検出する分には必要十分な情報であるが，類似シーンを検出する上では，コンテンツの特性を配慮した特徴情報の選択が必要と考えられる．

2.4 基本的なアイデア

スポーツ映像にはシーン特有のカメラワークが存在することが多い．例えば野球中継におけるホームランシーンでは，バッターが打ったボールをカメラで追い，ボールが落下したスタンドをズームアップし，次にダイヤモンドを回るバッターを追う，といった典型的な一連のカメラワークが存在する．サッカー映像においても同様なことが言え，例えばコーナーキックのシーンでは，コーナーからゴール前へ蹴り込まれたボールを追いつつゴール付近をズームアップするといった典型的なカメラワークが存在する．つまり，スポーツ映像を対象とする場合，カメラワークの類似性に基づき同じ意味合いのシーンを的確に検出できる可能性が高い．そこで本研究では，サンプルシーンとターゲット映像の照合に利用する映像の特徴情報として，カメラワークの変化を表す時系列データの利用を考える．パン操作やズーム操作の度合いを表すカメラパラメータは，MPEG 符号化情報に含まれる動きベクトルに基づき計算可能である[7]．

一方，時間軸方向の伸縮を考慮できる照合方式としては，音声認識の分野で提案された連続 DP マッチング[6]が著名である．これは元来，実時間に入力される発話を辞書に登録されている音声情報と効率良く比較するためのものであるが，短い時系列データを長い時系列データから切り出した部分区間と比較する方式と一般化して捉えられるので，本研究におけるサンプルシーンとターゲット映像の部分区間との照合処理にも応用できる．そこで本研究では，MPEG 符号化映像から抽出したカメラワークの時系列データに対して連続 DP マッチングを適用することで類似シーンを精度良く検出する方式を提案する．

3. MPEG 符号化情報に基づく類似シーン検出法

3.1 特徴ベクトルの抽出

文献 7)の手法に基づき MPEG 符号化情報の動き補償ベクトルからカメラパラメータを推定し，サンプルシーンおよびターゲット映像のカメラワーク情報を抽出する．P ピクチャとしてエンコードされたフレームについて，動き補償符号化されたマクロブロックの中心画素の位置を(x, y)，そのマクロブロックに対する動きベクトルを(u, v)とするとき，(u, v)が定点カメラの操作に伴う背景の動き（グローバルモーション）を表すならば次式が成り立つ．

$$\begin{pmatrix} u \\ v \end{pmatrix} = G_z \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} G_x \\ G_y \end{pmatrix} \tag{1}$$

ここで G_x, G_y, G_z はカメラパラメータであり，それぞれ水平方向のパン操作，垂直方向のパン操作，ズーム操作の度合いを表す．フレームを構成するすべての動き補償符号化マクロブロックを対象に式(1)を最小 2 乗法により解くことでカメラパラメータを推定できるが，実際にはすべての動きベクトルがグローバルモーションを表すわけではないため推定誤差が生じる．これに対処するため，次式を満たす動きベクトル(u, v)のみを対象に最小 2 乗法を繰り返すこととする．

$$\sqrt{(u-u')^2+(v-v')^2} \le E \tag{2}$$

ここで，(u', v')は最小 2 乗法で求めたカメラパラメータに基づき推定した動きベクトル，E は閾値を表す定数である．つまり，(u', v')との差が閾値を越える(u, v)は被写体の動き（ローカルモーション）を表すベクトルと判断し，これらを除外した上で再度最小 2 乗法を適用する．これを，ローカルモーションを表すベクトルが検出されなくなるまで繰り返すことで，カメラパ

ラメータの推定誤差を低減する．

ローカルモーションと判断してカメラパラメータの推定から除外した動きベクトルは必ずしも被写体の動きを表してはおらず，動き補償エラーに伴い発生したベクトルを含む可能性が高い．しかし，ここではこれらを一括してカメラワーク以外の代表的な動きと捉え，カメラパラメータの推定から除外した動きベクトルの平均値(L_x, L_y)を特徴情報の要素として扱う．

MPEG 映像を構成する i 番目の P ピクチャから抽出できる特徴ベクトルを $f(i) = (G_x, G_y, G_z, L_x, L_y)$とするとき，サンプルシーン S およびターゲット映像 T から抽出した特徴ベクトル列 F_S および F_T は次のように表現できる．

$$F_S = f_S(1) \prec f_S(2) \prec \cdots \prec f_S(N_S) \quad (3)$$

$$F_T = f_T(1) \prec f_T(2) \prec \cdots \prec f_T(N_T) \quad (4)$$

ここで，N_S および N_T はそれぞれ S および T に含まれる P ピクチャの総数である．

3.2 特徴ベクトル列のマッチング

連続 DP マッチングにより F_S を F_T の部分区間と照合し，F_S の終点 $f_S(N_S)$ を F_T の各要素 $f_T(k)$ $(1 \leqq k \leqq N_T)$ に対応付けたときの類似度 $D(k)$ を次式により求める．

$$D(k) = \frac{1}{k - k' + N_S} g(k, N_S) \quad (5)$$

$$g(i,j) = \min \begin{pmatrix} g(i-1, j-1) + 2d(i,j) \\ g(i-1, j-2) + 3d(i,j) \\ g(i-2, j-1) + 3d(i,j) \end{pmatrix} \quad (6)$$

尚，漸化式(6)の境界条件は次のように与える．

$$g(0,j) = g(i,0) = \infty \quad (7)$$

$$g(1,j) = \infty \qquad (2 \leq j \leq N_S) \quad (8)$$

$$g(i,1) = 2d(i,1) \quad (1 \leq i \leq N_T) \quad (9)$$

k'は，漸化式(6)の計算が終了したとき，つまり，j が 1 となったときの i の値であり，F_S の始点 $f_S(1)$が F_T の要素 $f_T(k')$と照合されることになる．$d(i,j)$は $f_T(i)$と $f_S(j)$ のユークリッド距離であり，次式により求める．

$$d_{pan} = (G_x(i) - G_x(j))^2 + (G_y(i) - G_y(j))^2 \quad (10)$$

$$d_{zoom} = (G_z(i) - G_z(j))^2 \quad (11)$$

$$d_{local} = (L_x(i) - L_x(j))^2 + (L_y(i) - L_y(j))^2 \quad (12)$$

$$d(i,j) = \sqrt{\omega_1 d_{pan} + \omega_2 d_{zoom} + \omega_3 d_{local}} \quad (13)$$

ω_1，ω_2，および ω_3 により類似度算出におけるグローバルモーションとローカルモーションの影響の度合いを調整できる．

連続 DP マッチングを適用することで，F_S と F_T の特徴ベクトルは単純に並び順に照合されるのではなく，最も類似性を高くできる対応付けが式(6)により動的に決定されることになる．

3.3 類似シーンの選択

前節の照合処理により，$f_T(k')$に対応するフレームを始点，$f_T(k)$に対応するフレームを終点とするターゲット映像中のシーン $s(k)$に対する類似度 $D(k)$が得られる．$s(k)$のうち類似度が高い，つまり $D(k)$の値が小さいものをサンプルシーン S に対する類似シーンとして単純に選択すると，時区間が大幅にオーバーラップした冗長なシーンが多数検出されてしまうため，次の手順で冗長なものを排除する．

先ず，$s(k)$を $D(k)$に基づき類似度順にソートする．次に，ソート順に $s(i)$を選択しながら次式の判定を繰り返し行い，条件を満たさない $s(j)$を除外する．

$$\frac{overlap(s(i), s(j))}{length(s(j))} \leq O \ (1 \leq i \leq N_T, i < j) \quad (14)$$

ここで，*overlap* は引数に与えたシーンの重複する時区間の長さを返す関数であり，*length* はシーン長を返す関数である．これにより，類似度の高いシーンとオーバーラップする時区間の割合が閾値 O を越える類似度の低いシーンが検出結果から排除される．

このようなフィルタリングを施しても，検出結果は一般にかなり多くの不要シーンを含むので，実際には類似度が著しく低いシーンを閾値処理により排除する，あるいはアプリケーションが要求する数だけ類似度の高いものから順に選択するというような処理を施す必要がある．

4. 実験と評価

4.1 実験素材と評価尺度

実際のプロ野球中継の映像を用いて提案法の有効性を評価した．ターゲット映像として使用した映像は，約 1 時間 40 分の野球中継を MPEG-1 符号化したもの（画像サイズ 352×240，フレーム数 171,502，約 1 GB）であり，46,476 個の P ピクチャを含む（N_T = 46,476）．また，ターゲット映像からホームランと内野ゴロの実況シーンをそれぞれ 1 つずつ選び，これら 2 つを実験のサンプルシーンとした．ホームランシー

表1 実験サンプル
Table 1 Examined samples

サンプル名	ホームラン	内野ゴロ
画像サイズ	352 × 240	
長さ（容量）	18秒（3.3MB）	7秒（1.4MB）
Pピクチャ数（N_S）	148	63
類似シーン数	10	11

ンは，投手がボールを投げる場面からダイヤモンドを回るバッターを映す場面までの画像から構成される．一方，内野ゴロシーンは，投手がボールを投げる場面からバッターがアウトになる場面までの画像から構成される．実験に用いたサンプルシーンの詳細は表1の通りである．

評価の尺度には，情報検索の評価で一般的に用いられている適合率と再現率を採用する．検出結果を類似度順にソートした並びの n 番目に i 個目の正解シーンが出現するとき，この出現位置に対応する適合率 $P(i)$ と再現率 $R(i)$は次式で求められる．

$$P(i) = \frac{i}{n} \tag{15}$$

$$R(i) = \frac{i}{N_c} \tag{16}$$

ここで，N_c は正解シーンの総数である．各正解シーンについて $P(i)$と $R(i)$を算出し，再現率を横軸，適合率を縦軸とした適合率－再現率グラフを得る．

適合率は再現率が高くなるにつれて劣化するのが一般的なので，適合率－再現率グラフは一般に右下がりのカーブとなる．適合率－再現率グラフに基づく評価ではグラフの下がり具合を性能の指標とする場合が多いが，本研究ではこの指標が最適とは言えない．本研究における類似シーン検出の目的は，タグ付け処理の効率化にある．そのためには，検出結果からすべての正解シーンを発見する上での手間を少なく抑えることが重要であり，この手間の度合いは再現率が 1.0 のときの適合率（以降，最終適合率とよぶ）で表現される．例えば，10 件の正解シーンを検出するとき，最終適合率が 0.5 であれば高々20 件の検出シーンをブラウジングするだけですべての正解シーンを発見できるのに対し，最終適合率が 0.1 であると 100 件ものシーンをブラウジングしなければならなくなる．以降の実験では，評価の一般性を考慮して適合率－再現率グラフを実験結果として示すが，上述の理由により本研究では最終適合率が特に重要であるため，主に最終適合率を指標として性能に関する議論を行う．

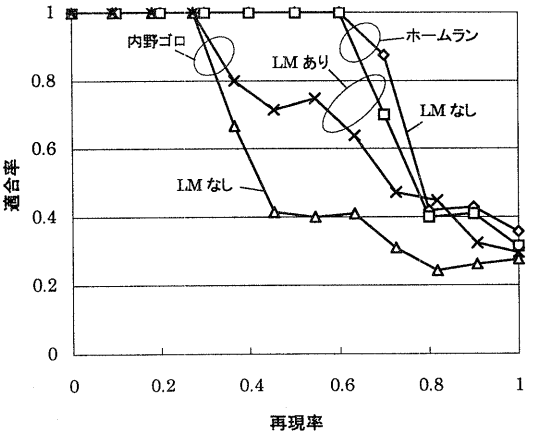


図2 提案法の基本性能
Fig. 2 Basic performance of proposed method

4.2 予備実験

初めに，提案方式の基本性能を評価する．式(13)において d_{pan}，d_{zoom}，d_{local} が同等に寄与するように ω_1，ω_2，ω_3 の値を設定した場合，および ω_3 を 0 としてローカルモーション（LM）を考慮しない場合について適合率と再現率を求めた．冗長なシーンを排除するための条件式(14)における閾値 O の値は 0.2 とした．結果を図2に示す．

内野ゴロシーンの検出では，ローカルモーションを考慮した場合としない場合とでグラフに大きな違いが生じており，ローカルモーションに関する特徴情報を照合処理に利用する本手法の有効性を示している．一方，ホームランシーンの検出では同様の効果が現れておらず，むしろローカルモーションを考慮しない方が，最終適合率が若干良好となっている．つまり，ローカルモーションの寄与の度合はサンプルシーン依存であり，検出したいシーンの特性に応じて適宜適切な重み付けを行う必要があることが分かる．

以上の結果を踏まえ，以降の実験では，ホームランシーンの検出においては式(13)の ω_3 を 0 としてローカルモーションを考慮しない重み付けとし，一方，内野ゴロシーンの検出においては d_{pan}，d_{zoom}，d_{local} が同等に寄与するように ω_1，ω_2，ω_3 の値を設定して適合率と再現率を求めることとする．

4.3 カメラワーク情報を利用する効果

提案法の主たるアイデアは，スポーツ映像中の類似シーンに共通するカメラワークの存在に着目した点である．ここでは，従来法で一般に用いられている色情

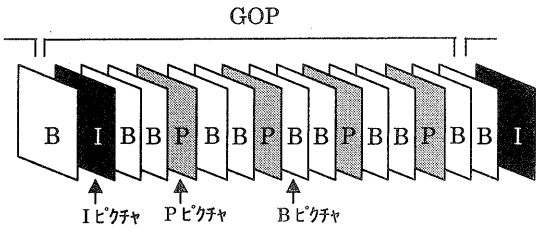


図3 MPEG符号化映像の基本構造
Fig. 3 Basic structure of MPEG encoded video

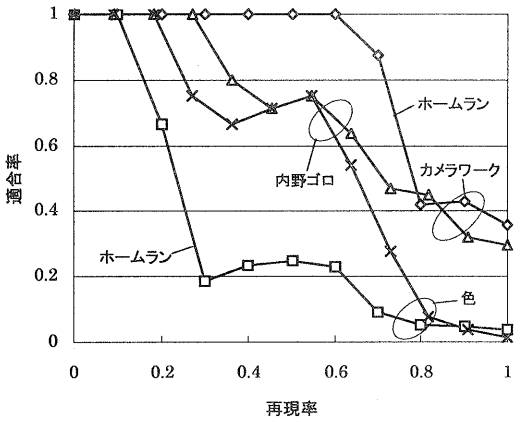


図4 カメラワーク情報を用いる効果
Fig. 4 Effect of using camera work feature

報を利用した場合との性能比較を通して，カメラワーク情報を利用する効果を定量的に評価する．

MPEG符号化情報では，Iピクチャに含まれるマクロブロックの離散コサイン（DCT）係数に基づき色情報を抽出できる．そこで，DCT係数の直流成分に基づき，フレーム画像の平均輝度（Y），青色に対する平均色差（C_b），および赤色に対する平均色差（C_r）を求め，Iピクチャ単位に色情報に関する特徴ベクトルを生成した．実験に用いたMPEG符号化映像の基本構造が図3のようなGOP（Group Of Pictures）を単位とすることから，Pピクチャ毎に特徴ベクトルを抽出する場合に比べ，特徴ベクトル数は約4分の1となる．i番目のIピクチャから抽出できる特徴ベクトルを$f(i)=(Y, C_b, C_r)$として前述の式(3)および式(4)の特徴ベクトル列F_SおよびF_Tを構成し，3.で述べた手法により類似シーンを検出した結果，図4のようになった．尚，$f_T(i)$と$f_S(j)$のユークリッド距離$d(i, j)$の算出には，式(10)～(13)の代わりに次式を用いた．

$$d_Y = (Y(i) - Y(j))^2 \tag{17}$$

$$d_{Cb} = (C_b(i) - C_b(j))^2 \tag{18}$$

$$d_{Cr} = (C_r(i) - C_r(j))^2 \tag{19}$$

$$d(i, j) = \sqrt{\omega_4 d_Y + \omega_5 d_{Cb} + \omega_6 d_{Cr}} \tag{20}$$

ω_4，ω_5，ω_6の値はd_Y，d_{Cb}，d_{Cr}が同等に寄与するように設定した．

このように，ホームランと内野ゴロどちらのサンプルシーンについても，色情報に基づき照合処理を行った場合の最終適合率はカメラワーク情報を用いた場合に比べ著しく劣っている．つまり，提案法は，色情報を利用する従来法に比べ，タグ付け処理の効率化に適した方式と言える．

4.4 連続DPマッチングを適用する効果

提案法では，類似シーン毎の長さの違いを考慮した照合処理を実現する目的で，特徴ベクトル列を時間軸方向に伸縮させながら最適な照合パターンを選択できる連続DPマッチングを採用した．ここでは，連続DPマッチングを適用する効果を，特徴ベクトル列を単純に並び順に照合する場合との比較を通して定量的に評価する．

連続DPマッチングでは，F_SとF_Tの特徴ベクトルの最適な対応付けを式(6)により動的に決定することで，特徴ベクトル列を時間軸方向に伸縮させた照合処理を実現している．つまり，式(6)を次式に置き換えれば，F_SとF_Tの特徴ベクトルを単純に並び順に照合する方式（単純照合）に等価となる．

$$g(i, j) = g(i-1, j-1) + 2d(i, j) \tag{21}$$

式(6)を式(21)に置き換えて性能を測定した結果，図5のようになった．このように，連続DPマッチングは単純照合より良好な最終適合率を提供でき，タグ付け処理の効率化に効果的であることが分かる．内野ゴロシーンの検出ではあまり大きな効果が現れていないが，これは内野ゴロシーンのシーン長のばらつきが小さかったためと考えられる．一方，ホームランシーンの検出では，最終適合率に大きな違いが生じている．これは，検出すべきホームランの類似シーンの中に，サンプルシーンと場面構成が異なるシーンが含まれていたことに起因する．サンプルに用いたホームランシーンは投手がボールを投げる場面から開始されるのに対し，この例外的なシーンはバッターが打つ場面から開始され，投手がボールを投げる場面が存在しない（投手が投げる場面が放送されていない）．連続DPマッチングによれば，このような場面構成の違いにも柔軟

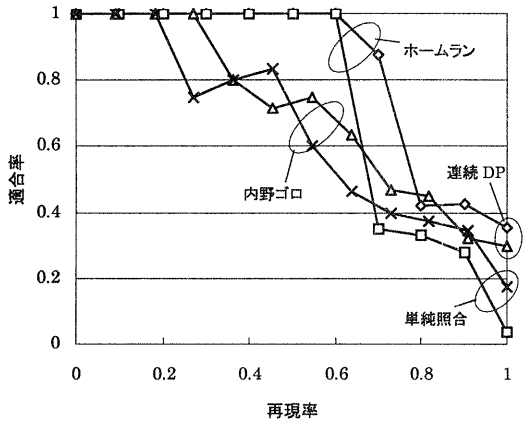


図5 連続 DP マッチングの効果
Fig. 5 Effect of using continuous DP matching

に対応できることが分かる.

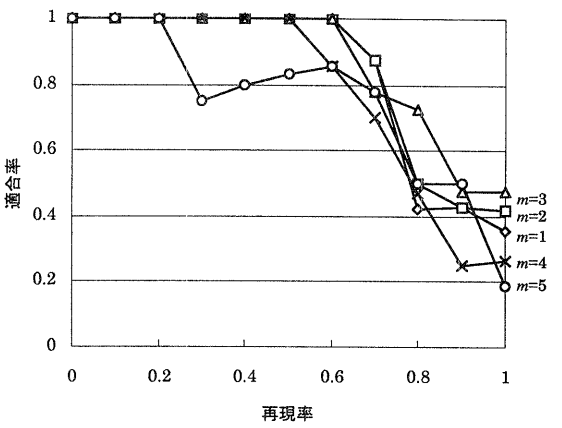
4.5 特徴ベクトル列の圧縮法

これまでの評価では，検出精度という観点で提案法を従来法と比較し，その有効性を明らかにした．ここでは，検出速度の観点から提案法について考える.

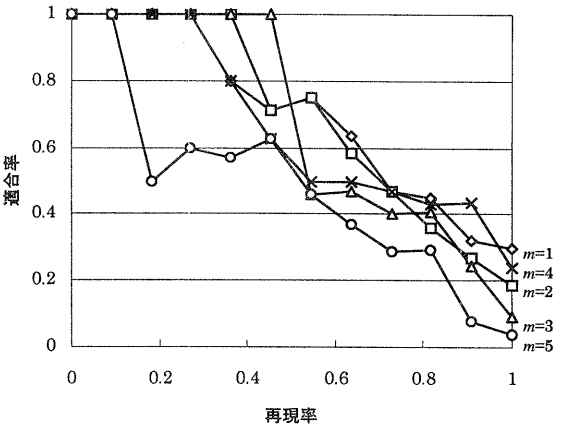
提案法では，連続DPマッチングの性質から $O(N_S*N_T)$ の検出処理時間を要する．例えば 4.2 で述べたホームランシーンの検出を PentiumII 400MHz のプロセッサで実行した場合，およそ 25 秒もの時間を要することになり，これは実用的な検出速度とはいえない．検出速度向上のためには映像情報から抽出する特徴ベクトル数を削減する必要があるが，F_S および F_T から特徴ベクトルを単純に間引くと検出精度の劣化につながる可能性がある.

特徴ベクトルの間引きによる影響を明確化するため，特徴ベクトルの生成対象とするPピクチャのサンプリング間隔 m を変化させて適合率と再現率を求めた．結果を図 6 に示す．尚，検出速度は m の値に応じて図7のようにほぼ理論値通りに向上した．図7の縦軸は，m が1，つまり，すべてのPピクチャを対象に特徴ベクトル列を形成したのときの検出処理時間に対する比を表す.

このように，サンプリング間隔を大きくすると，検出処理時間は向上するものの，適合率は大幅に劣化してしまう．特に，内野ゴロシーンのようにもともとのシーン長が短い場合，特徴ベクトルの間引き方により特徴情報の欠落の度合いが大幅に変化するため，サンプリング間隔による適合率の変動が激しくなってしまう.



(a) ホームランシーンの検出



(b) 内野ゴロシーンの検出

図6 特徴ベクトルの間引きの影響
Fig. 6 Effect of feature vector curtailment

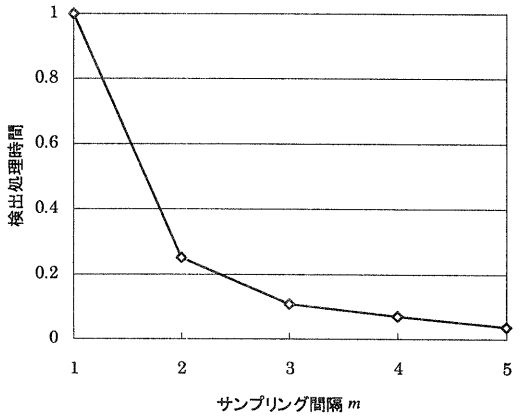
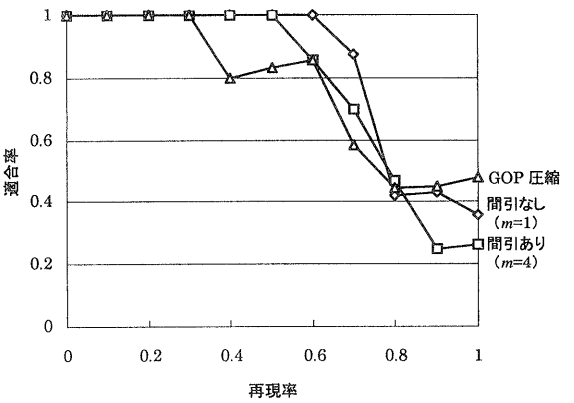


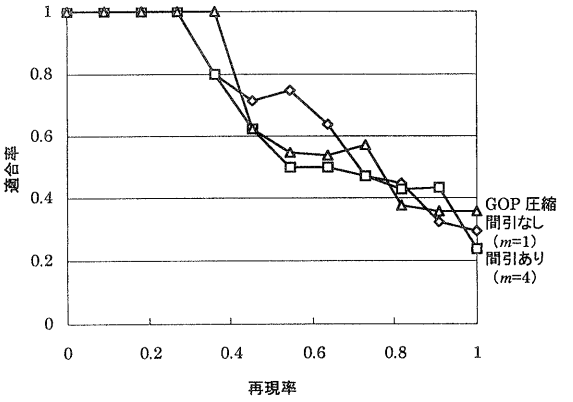
図7 特徴ベクトルの間引きに伴う検出処理時間の変化
Fig. 7 Detection time with varying sampling interval of feature vectors

特徴ベクトル数を削減するもう1つの手法として，連続するいくつかの特徴ベクトル毎に平均ベクトルを求め，得られた平均ベクトル列を照合処理に用いる方法が考えられる．間引きによる手法と異なり，全ての特徴情報を圧縮した形で照合処理に利用できるため，検出精度の劣化抑制が期待できる．

GOP 単位に特徴ベクトルの平均値を求める方法で特徴ベクトル列の圧縮を図ったときの性能を図8に示す．GOP には4つの P ピクチャが含まれるので，特徴ベクトル数はサンプリング間隔を4として間引きを行ったときと同等となる．このように，GOP 単位での圧縮法によれば，すべての P ピクチャから抽出した特徴ベクトルを対象に照合処理を行ったときと同等の最終適合率を提供できると共に，検出処理時間を 16 分の 1 に短縮できることが分かる．25 秒を要していたホームランシーンの照合処理は約 1.6 秒で完了することになり，これは十分実用的な値と考えられる．



(a) ホームランシーンの検出



(b) 内野ゴロシーンの検出

図8 特徴ベクトル列の圧縮効果

Fig. 8 Effect of feature vector compression

5. おわりに

本論文では，MPEG 符号化情報から直接的に求められる映像の特徴情報を用いて，映像に含まれる類似シーンを精度良く検出する方式を提案した．提案法は，主にスポーツ映像へのタグ付け処理の効率化を狙いとし，スポーツ映像の類似シーンに共通するカメラワークの存在に着目してシーン検出を行うことを特長とする．また，連続 DP マッチングをカメラワーク情報の照合処理に適用することで，類似シーン毎のシーン長の違いに柔軟に対応できる．

実際の野球中継の映像を用いて実験した結果，カメラワーク情報に基づき類似シーンを検出する手法は，従来一般に用いられている色情報を用いる手法に比べて良好な適合率と再現率を提供できることが明らかとなった．また，特徴ベクトル列を単純に並び順に照合する手法との比較を通して，連続 DP マッチングを適用する効果を明確化した．更に，照合処理時間の短縮化法について検討し，特徴ベクトル列を GOP 単位に圧縮する手法により，類似シーンの検出精度を劣化させることなく照合処理時間を実用的な値まで短縮できることを示した．

提案法の主たるアイデアは類似シーンに共通するカメラワークの存在に着目した点であり，この規則が当てはまらないケースにはうまく対応できない．例えば，野球中継の映像では，今回の実験で用いた典型的なホームランシーンのほか，スタンド側からボールを追ったホームランシーンが含まれる場合もある．この例外的なシーンも類似シーンの定義からすれば検出対象となるが，これを典型的なホームランシーンとの比較から検出するのは明らかに困難である．この問題に対しては，シーンのバリエーションを考慮したシーン辞書を作り，複数のサンプルシーンを組み合わせて1つの類似シーン集合を求めるなどの方法で対処できると考えられる．また，今回の実験を通して色情報のみでは十分な検出精度を得られないことが分かったが，カメラワーク情報と色情報を複合的に用いることで，カメラワーク情報単独の場合より検出精度を高められる可能性がある．例えば，野球中継の途中に現れる CM シーンにホームランシーンと同じようなカメラワークが用いられている場合，カメラワーク情報だけで照合処理を行うとその CM シーンが誤検出されてしまうが，色情報を組み合わせれば，両者の色の違いによりこの誤検出は防止できると考えられる．今後，これら新しいアイデアついて実証を進める予定である．

参考文献

1) 桑野秀豪, 倉掛正治, 小高和己: 映像データ検索のためのテロップ文字抽出法, 信学技報, PRMU96-98, pp.39-46 (1996).
2) 宮森恒ほか: シーン中の短時間動作記述を用いた映像内容検索方式の提案, 画像の認識・理解シンポジウム予稿集 I, pp.75-80 (1998).
3) 有木康雄: 音声単語スポッティングに基づくテレビニュース記事の自動分類, 電子情報通信学会論文誌, D-I, Vol.J82-D-I, No.1, pp.223-233 (1999).
4) 柏野邦夫, ガビンスミス, 村瀬洋: ヒストグラム特徴を用いた音響信号の高速探索法—時系列アクティブ探索法—, 電子情報通信学会論文誌, D-II, Vol.J82-D-II, No.9, pp.1365-1373 (1999).
5) 長坂晃朗, 宮武孝文: 時系列フレーム特徴の圧縮符号化に基づく映像シーンの高速分類手法, 電子情報通信学会論文誌, D-II, Vol.J81-D-II, No.8, pp.1831-1837 (1998).
6) 新美康永: 音声認識, 情報科学講座 E・19・3, 共立出版(株) (1980).
7) J. Meng and S. Chang: "CVEPS – A Compressed Video Editing and Parsing System", *Proc. ACM Multimedia'96*, pp.43-53 (1996).
8) 佐藤隆ほか: MPEG 符号化映像からの高速テロップ領域検出法, 電子情報通信学会論文誌, D-II, Vol.J81-D-II, No.8, pp.1847-1855 (1998).
9) 楊楊, 中野愼夫: MPEG2 符号化動画像データからのカメラワーク抽出法の検討, 信学技報, IE98-175, pp.39-44 (1999).
10) 粕谷英司ほか: 圧縮映像中のパラメータを利用した高速照合とその検索方式の提案, 情処研報, AVM17-5, pp.25-32 (1997).
11) 新倉康巨ほか: MPEG 符号化映像ショットチェンジ検出のための動き補償解析ハイブリッド法の提案, 電子情報通信学会論文誌, D-II, Vol.J81-D-II, No.8, pp.1838-1846 (1998).

(平成 11 年 12 月 20 日受付)
(平成 12 年 3 月 1 日採録)

(担当編集委員 加藤 俊一)

片岡 良治 (正会員)

1985 年千葉大学工学部電子工学科卒. 1987 年同大学院電子工学専攻修士課程修了. 同年, 日本電信電話株式会社入社. 以来, トランザクションの並行処理制御方式の研究, マルチメディア情報システムの研究に従事. 電子情報通信学会会員.

遠藤 斉 (正会員)

1995 年東京工業大学理学部物理学科卒. 1997 年東京大学大学院総合文化研究科広域科学専攻修士課程修了. 同年, 日本電信電話株式会社入社. 以来, マルチメディア情報システムの研究に従事.