データベースシステム 120－11
（2000．1．24）

動画像の特徴量を用いた意味的構造の自動検出

谷澤和昭† 上原邦昭††

できごとや情景といった意味的情報に基づいて動画像データにアクセスするためには，内容に基づいたインデックス付けが必要である．インデックス付けで重要な問題は，動画像データの中からある決まった意味的情報を表している動画像区間をどのようにして特定するかという点である．本稿では，意味的なまとまりを持つ動画像区間を連続した動画像ショット列として定義し，定義に基づいてショットに分割された動画像データから情報として特徴量を自動的に抽出し，音声言語の確率モデルである N-gram モデルを導入して動画像区間を発見する手法を提案する．また，N-gram モデルの応用として，部分的な動画像から原形となった動画像の全体を特定する実験について考察する．

Automatic Detection of the Semantic Structure from Video by Using N-gram Model

KAZUAKI TANIZAWA † and KUNIAKI UEHARA††

For indexing video data based on contents, it is necessary to access video data based on semantic information like a certain event and scene. The most important process in indexing is to determine a paticular appropriate video interval with specific semantic information. This is called semantic structure. In this paper, we propose an algorithm for discovering semantic structures of video data. This algorithm is to discover semantic video intervals as consecutive sequences of video shots. We show the system to extract the amount of characteristics of video data from video shots, and to retrieve paticular video interval from shot information by using a probability model which is proposed in the field of speech recognition.

1. はじめに

今日，コンピュータの処理性能や記憶容量の飛躍的な向上とデジタル処理技術の進歩により，これまで様々な媒体上に蓄積されていたデータがデジタル化され，コンピュータ上で自由に処理し，加工できるようになってきた．それにともない，画像や音声などのマルチメディアデータの量も飛躍的に増大し，これらをデータベース化したいという要求が徐々に高まってきている．

動画像データは，今日，その利用が最も期待されているマルチメディアデータの１つである．動画像データの特徴は，画像や音声といった物理的な情報から，それを用いて表現されるできごとや情景といった意味的な情報に至るまで，様々な種類の情報が単一のデータ中に混在していることである．そのため，動画像データをデータベース化する際には，アクセスしたい情報に基づいてインデックス付けを行う必要がある．これまで，動画像データの画像情報や文字認識，音声認識によって得られる情報に基づいたインデックス付けにより，デジタルライブラリーシステム[2)3)]やNews On Demand システム[4)5)]などが開発されている．

一方，動画像データを意味的情報，すなわち動画像で表現されるできごとや情景などに基づいてインデックス付けを行う場合には，意味的情報を表現した時間区間の動画像，すなわち動画像区間を特定し，その内容を記述することが必要となる．このような動画像区間の特定には，専門的な知識や経験が必要なため，インデックス付けを行う記述者が人手で動画像区間の特定を行ってきた．しかしながら，あらゆる意味的情報に対する動画像区間を記述者が逐一定義するのは大変な作業であり，かなりの時間を必要としている．

以上の問題を解決するため，ショットの内容記述を用いて，動画像データの中から何らかの意味的なまとまりを表している動画像区間を発見する研究[8)]が行われている．この手法を用いれば，従来，すべて人手

† 神戸大学大学院自然科学研究科
Graduate School of Science and Technology, Kobe University
†† 神戸大学都市安全研究センター
Research Center for Urban Safety and Security, Kobe University

で行っていたインデックス付け作業の負担を大幅に軽減できるが，ショットに対する内容記述は事前に記述者が行なわなければならず，膨大な量の動画像に対する負担は大きいままである．

記述者の負担を減らすためには，ショットの内容記述を自動化する必要がある．ショットに関する情報として，ショットの中にあらわれる登場人物や背景，単純なアクションなどは，今日の画像処理技術や音声処理技術によってある程度まで認識可能である[9)10)]．これらの認識技術を用いれば，ショットの内容を自動的に抽出し，記述者の負担を大幅に軽減できると考えられる．しかしながら，これらの技術で動画像の内容を抽出するためには，高度な画像処理技術や認識技術が必要であり，部分的には人間の認識能力に頼らざるを得ないのが現実である．

一方，本研究では，ショットの情報として動画像に含まれる画像の色情報を用いている．色情報は画像処理技術を用いて自動的に取り出すことができるため，いっさい人手を必要とせずにショットの情報を得られる．したがって，この情報を用いて動画像データの中から動画像区間を発見すれば，動画像データベースにおけるインデックス付け作業の支援システムとして利用できると考えられる．

2. 動画像の意味的構造の発見

動画像データは，様々な出来事や情景を表現できるように，その基本構成要素であるショットをつなぎ合わせて構成される．したがって，内容が大きく変化するショットを見つけ出すことができれば，シーンの境界点を特定できる．また，ショットの内容を比較し，あるできごとや情景を表すために構成されたショット列を見つけ出せれば，特定の意味的情報をもつ動画像区間を特定できる．

動画像の意味的構造とは，ショットの内容に基づいて発見できる意味的にまとまった動画像区間やシーンである．ショットの内容を表す情報として，本研究ではショット間の色度数ヒストグラム差分を用いている．ショット間の色度数ヒストグラム差分の計算には χ^2 検定法[1)]を用いており，2 つのショット s_1，s_2 に対して式（1）による評価を行っている．

$$d_1 = D(s_1, s_2) = \sum_j \frac{\{H_c(s_1, j) - H_c(s_2, j)\}^2}{H_c(s_1, j)} \quad (1)$$

ここで，$H_c(s, j)$ は色度数ヒストグラムのことで，ショット s の先頭フレームにおいて色レベル j を持つ画素の数を示している．

2.1 確率モデル

得られたヒストグラム差分の列から意味的にまとまった意味的構造を発見するために，音声言語の生成確率モデルである N-gram モデル[6)7)]を導入する．N-gram モデルは音声認識の分野で盛んに用いられている言語モデルであり，ある単語をもとに次の単語を予測し，認識率の向上や計算時間の削減を目指すものである．例えば，n 単語からなる単語列 $w_1, w_2, \cdots, w_n$ (以下，w_1^n) が与えられたとき，単語 w_n の生成確率は N-gram モデルでは式（2）で与えられる．

$$P(w_n|w_1^{n-1}) = P(w_n|w_{n-N+1}^{n-1}) \quad (2)$$

すなわち，単語 w_n の生成確率 $P(w_n|w_1^{n-1})$ は直前の $N-1$ 個の単語にのみ依存し，生成確率が高い場合は単語 w_n の推定が妥当であると考えられる．

動画像においては，映像編集においてショットつなぎの技法といったものが存在しており，まとまったできごとや情景を表現するためにある決まったパターンが用いられている．これを文章における文法とし，映像パターンを単語列と考えると，動画像の N-gram モデルにおいて，あるできごとや情景を表すまとまった区間では N-gram の値が高くなると考えられる．一方，N-gram の値が急激に大きくなったり小さくなったりしている部分は，そこから新しいシーンに切り替わったり，重要な場面であると考えられる．そのため，このような状態が見られるショット間はシーンの境界として検出できると考えられる．

しかしながら，N-gram モデルは静的なモデルであり，モデルのパラメータは学習データに依存し，モデルが適用されるデータとは無関係に決定される．そのため，N-gram モデルにないできごとや情景を表す区間は発見できない可能性がある．したがって，モデルのパラメータを動的に変化させるために，キャッシュと呼ばれる直前の M 単語の単語分布を用いるキャッシュモデルを適用することを考える．

キャッシュモデルでは，単語 W_n の生成確率は直前の M 単語 w_{n-M}^{n-1} 中に w_n と同じ単語がどれだけ含まれていたかを考慮して決定される．すなわち，直前に使われた単語は再び使われやすいという単語の局所的な性質を反映したモデルとなっている．キャッシュに基づく単語 w_n の確率 $P_c(w_n)$ は，式（3）で与えられる．

$$P_c(w_n) = \sum_{m=1}^{M} a_m \delta(w_n, w_{n-m}) \qquad (3)$$

ここで，a_m は単語位置に対する重み係数で，$0 \leq a_n \leq 1$ かつ $\sum_m a_m = 1$ を満たすよう決定される．$\delta(\cdot)$ はクロネッカーの δ 係数であり，引数が等しいときは 1，それ以外は 0 を与える関数である．

キャッシュモデルを動画像に適用し，動的にモデルのパラメータを決定すれば，直前の数ショットから次に現れるショットの妥当性が検証でき，できごとや情景を表すまとまった区間を発見できると考えられる．また，N-gram モデルの場合と同様に，キャッシュモデルの値が急激に大きくなったり小さくなったりしている部分のショット間ではシーンの境界が存在すると考えられる．

2.2 意味的構造の発見

2.2.1 キャッシュモデルを用いた動画像区間の発見とシーン境界の検出

2.1 で述べたキャッシュモデルを動画像の色度数ヒストグラム差分列 $d_1, d_2, \cdots, d_n$ に適用すると，式 (3) は，

$$P_c(d_n) = \sum_{m=1}^{M} a_m \delta(d_n, d_{n-m}) \qquad (4)$$

と置き換えられるが，色度数ヒストグラム差分の値 d_i は単語と違って，完全に一致することは非常に少ない．そこで，クロネッカーの δ 係数をショット間のヒストグラムの類似度で置き換える．すなわち，式 (4) における δ を，

$$\delta(d_i, d_j) = similarity(i, j) \qquad (5)$$

で置き換える．ここで，$similarity(i, j)$ はショット i，j の類似度を計算する類似度関数である．

ショットの内容記述から動画像区間を発見する手法[8]では，意味的構造の中で定義される動画像区間を，以下の 3 種類に分類している．

(1) **Unchanged**
登場人物や背景などが動画像区間を通して変化しないことによって，それらを強調する．

(2) **Gradually Changing**
登場人物や背景などがショットごとに徐々に変化することによって，あるアクションや情景を表現する．

(3) **Multiplexing**
個々の登場人物や背景などがショットごとに交互にあらわれることによって，各々の繰り返しの中で表現されているできごとや情景を互いに関連づける．

これに対し，本稿では，式 (4) の重み係数 a_m として，以下の 2 種類を用いて，意味的構造の中で定義される動画像区間を分類している．

(1) 変化量が一定の区間 ($a_m = \frac{1}{m}$)
キャッシュ内のすべての差分に同様の重みを与える．

(2) 変化量が徐々に変わる区間
($a_m = \frac{k}{m}$, $\sum_m \frac{k}{m} = 1$)
時間的に近い差分値ほどより重要であると考え，重みを大きくする．

式 (4) において (1),(2) のそれぞれの重み係数を適用し，得られる確率 $P_c(d_n)$ が閾値 θ を満たせば，色度数ヒストグラム差分列 $d_{n-M}, d_{n-M+1}, \cdots, d_n$ は意味的構造となる動画像区間として検出される．

動画像区間の発見は，以下の手順で行われる．

(1) すべてのショット間の色度数ヒストグラム差分に対して，式 (4) を適用して確率を計算する．類似度関数 $similarity(i, j)$ は，ヒストグラム差分値の差の最大値を 0，最小値を 1 として正規化する．

(2) 類似度閾値 θ を与えて，確率の妥当性を判断する．確率が閾値を満たせば，キャッシュモデルとして用いたヒストグラム差分列を動画像区間として検出する．

また，2.1 で述べたように，キャッシュモデルの確率が急激に大きくなったり小さくなったりしている部分を，新しいシーンや重要な場面の始まりであると考える．これを先に述べた 2 種類のキャッシュモデルへの重みそれぞれに対して求め，シーンの境界として検出する．

2.2.2 N-gram モデルを用いたシーン境界の検出

2.1 で説明したように，キャッシュモデルは入力データごとにモデルのパラメータを変化させることができるが，入力データに特化したモデルとなるため汎用性はない．そこで，汎用性を持たせるため N-gram モデルを導入する．N-gram モデルでは，大量の学習データをもとにヒストグラム差分値の妥当性を検証している．したがって，キャッシュモデルのように動画像ごとに正規化を行うと，モデル内での動画像データごとの互換性が失われる．このため，ヒストグラム差分値を離散化して，多数の動画像に対して互換性のあるモ

デルを構築する．

N-gram モデルでは，通常 $N=2$（bigram），または $N=3$（trigram）が用いられるが，本稿では 2-gram（bigram）モデルを適用する．これは，対象データがショット間の色度数ヒストグラム差分値であり，前後 2 つのショットの情報を持っているため，bigram モデルで前後 3 ショットの情報が得られると考えられるためである．以上のことから，離散化色度数ヒストグラム差分列 $d'_1, d'_2, \cdots, d'_n$ における d'_n の生成確率は，式（6）で与えられる．

$$P(d'_n|d'_{n-2}, d'_{n-1}) = P(d'_n|d'_{n-1}) \qquad (6)$$

シーン境界の検出は以下の手順で行われる．

（1） 2-gram モデルの構築
- （a） ショット間の色度数ヒストグラム差分列 $d_1, d_2, \cdots, d_n$ に対して離散化を行い，$d''_1, d''_2, \cdots, d''_n$ を求める．
- （b） 離散化データをもとに，離散値を単語とする 2-gram モデルを構築する．

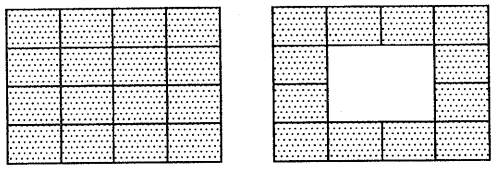
（2） モデルに対する対象データの評価
- （a） ショット間の色度数ヒストグラム差分列 $d_1, d_2, \cdots, d_m$ に対して離散化を行い，$d''_1, d''_2, \cdots, d''_m$ を求める．
- （b） $d''_1, d''_2, \cdots, d''_m$ の隣接するショット d''_i，d''_{i+1}（$1 \leq i \leq m$）すべてにおける 2-gram モデルに対する適切度（relevance）を求める．
- （c） 適切度の極端に大きくなったり小さくなっている点をシーンの境界として検出する．

2.2.1 と同様に，2-gram の値が急激に変化する部分を，新しいシーンや重要な場面の始まりとして検出する．

また，2.2.1 で述べた動画像区間内でシーンの境界点が検出される場合がある．これは区間内で瞬間的に特別なできごとや情景が挿入されている場合などが考えられる．このような場合はシーンは分割されていないと考え，シーンのマージングを行っている．

2.2.3 シーン境界の種類の定義

シーンの移り変わりは，登場人物と舞台がどちらも変わってシーンが切り替わる場合と，街のシーンから家の中のシーンへと移るが登場人物は変わらないといった背景の変化のみが変化する場合が考えられる．このため，画面全体の情報を抽出してシーン検出を行う以外に，画面の外側部分の情報を背景として抽出し，背景が変化する点でのシーンの検出を行っている（図 1）．



(a) Whole block　　(b) Outside block

図1　シーン境界検出に用いる画面領域

図 1(a) ではすべてのブロックの情報を抽出しているが，図 1(b) では画面中央のブロックの情報は抽出していない．したがって，図 1(b) では背景部分の画像情報を抽出したことになり，背景の変化をとらえられる．

3. 実験と評価

実際の動画像を用いて，3.1 で述べた意味的構造発見の評価実験を行い本手法の有効性を検証した．実験には，約 2 時間のアニメーション映画 2 本の一部分を用いている．動画像形式は，いずれも日本 SGI 社の動画像フォーマットである SGI movie 形式を用い，秒間 30 フレーム，横 160 ピクセル，縦 121 ピクセルである．ショットへの分割はフレーム画像色分散の χ^2 検定に基づくカット検出プログラム[14] を用いている．

3.1 動画像区間の発見に関する評価

意味的構造として定義した動画像区間の動画像全体に対する汎用性，および発見精度を評価するため，約 8 分間（15000 フレーム），100 ショットの動画像を対象に発見アルゴリズムによって検出された動画像区間が動画像データ全体に占める割合と，その区間が正しいかどうかについて検証した．まず，類似度閾値の変化に対する動画像区間の割合と精度を検証した．なおキャッシュサイズは 4 としている．実験結果を表 1 に，結果から得られた類似度閾値の変化に対する動画像区間の割合と精度を図 2 に示す．

表1　類似度閾値に対する動画像区間の変化

Cache size	4	4	4
Threshold	0.5	0.6	0.7
Interval detected	4	11	7
Correct interval	0	9	6

表 1 において，Cache size はキャッシュモデルにおけるキャッシュの大きさ，Threshold は類似度閾値，

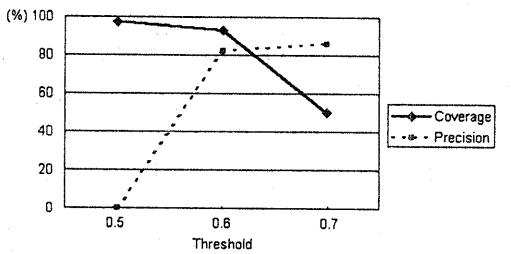


図2 類似度閾値に対する動画像区間のカバー率と適合率の変化

Interval detected は発見された動画像区間の数，Correct interval は発見された動画像区間のうち正しい動画像区間の数を示す．表 1 を見ると，閾値 0.7 の場合の動画像区間数が 0.6 の場合より少なくなっているが，これは 0.7 の場合は動画像区間が長くなり，多くのショットをまとめてしまったためである．逆に 0.5 の場合は閾値が厳しくなったため，0.6 の場合より動画像区間数が少なくなったと考えられる．

一方，図 2 の Coverage は，発見された動画像区間が動画像全体に対して占める割合（ショット数に基づく）を示している．Precision は動画像区間の適合率であり，式（7）で求められる．

$$(\text{適合率}) = \frac{(\text{正しい動画像区間の数})}{(\text{発見された動画像区間の数})} \quad (7)$$

図 2 において，類似度閾値が 0.5 の場合の適合率が 0% であるのは，類似判定の基準が低いため，本来は類似していない部分を誤判定して，動画像区間として検出したためである．この場合は，動画像区間の数が極端に少なくなり，区間の長さ自体も非常に長くなる．したがって，閾値の設定が誤っていると容易に判断でき，閾値の値を再設定すればよい．逆に，閾値を高くした場合は類似判定の基準が厳しくなるため，動画像区間が全体に占める割合は小さくなるが，動画像区間の精度は向上するという利点がある．また，類似度閾値が 0.6，0.7 の場合は 80% 以上の高い適合率を示しており，閾値の設定が適切であったと考えられる．

次に，キャッシュサイズの変化に対する動画像区間の割合と精度を検証した．なお閾値は 0.6 としている．実験結果を表 2 に示す．

表2 キャッシュサイズに対する動画像区間の変化

Cache size	3	4	5	6
Threshold	0.6	0.6	0.6	0.6
Interval detected	12	11	10	10
Correct interval	9	9	7	5

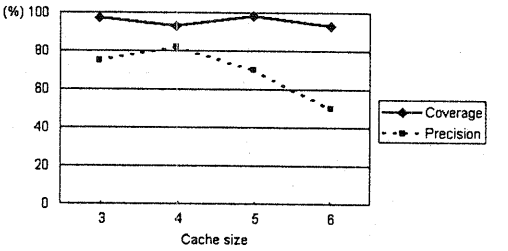


図3 キャッシュサイズに対する動画像区間のカバー率と適合率の変化

表 2 を見ると，キャッシュサイズが変化しても，検出する動画像区間の数はほとんど変化していない．これは，類似度閾値を緩めに設定していたために，動画像区間の発見がキャッシュサイズの変化にそれほど影響を及ぼさなかったと考えられる．実際に動画像区間を見ると，キャッシュサイズが大きくなるにつれて長い動画像区間が多くあらわれるようになり，精度に影響を及ぼしている．実験結果から得られたキャッシュサイズの変化に対する動画像区間の割合と精度を図 3 に示す．

図 3 より，キャッシュサイズが 6 のときに適合率が大きく低下しているのがわかる．本稿のアルゴリズムでは，キャッシュモデルの値が閾値を満たした場合，キャッシュに含まれるショットすべてを動画像区間としている．すなわち，キャッシュサイズが動画像区間の（ショット数に基づく）最小単位となる．1 ショットあたりのおおよその時間は対象とする動画像によって異なるが，キャッシュサイズが大きくなるとキャッシュモデルによる最小の動画像区間の長さが実際の動画像区間の長さを大きく上回ることも起こりうる．そのため，キャッシュサイズが大きい場合は同じ区間に入るべきでないショットが含まれてしまい精度が落ちる可能性がある．

この問題を解決するためには，対象とする動画像に応じて適切なキャッシュサイズを設定する必要がある．なお，キャッシュサイズが 3 から 5 の場合には 70% 以上の高い適合率を示しており，キャッシュサイズが適切であったことがわかる．

3.2 シーン境界の検出に関する評価

キャッシュモデル及び 2-gram モデルによるシーン境界の検出精度を評価するため，アルゴリズムによって検出されたシーン境界を検証した．対象とする動画像は，約 5 分間（10000 フレーム）の動画像からなる 2 種類（A，B）である．それぞれのショット数は，動

表3　確率モデルによるシーン境界の検出結果

	Movie A						Movie B					
	Cache only		Cache+Interval		Relevance		Cache only		Cache+Interval		Relevance	
	All	Out	All	Out	All	Out	All	Out	All	Out	All	Out
Boundary detected	32	32	17	17	15	16	28	28	13	15	12	18
Correct boundary	18	18	9	9	11	10	18	20	10	10	9	10
Precision (%)	56.3	56.3	52.9	52.9	73.3	62.5	64.3	71.4	76.9	66.7	75.0	55.6

画像 A が 66，動画像 B が 54 である．シーン境界の種類としては，次の 4 種類を用いている．

(1) 画面全体の情報に対してキャッシュモデルの値のみを用いたシーン境界
(2) 画面外側の情報に対してキャッシュモデルの値のみを用いたシーン境界
(3) 画面全体の情報に対してキャッシュモデルに動画像区間を組み合わせたシーン境界
(4) 画面外側の情報に対してキャッシュモデルに動画像区間を組み合わせたシーン境界
(5) 画面全体の情報に対して 2-gram モデルを用いたシーン境界
(6) 画面外側の情報に対して 2-gram モデルを用いたシーン境界

(2),(4),(6) については，図 1 のように画面を 16 (4 × 4) に分割し，外側の領域 12 個から情報を抽出した．キャッシュモデルにおけるキャッシュサイズは 4，類似度閾値は 0.6 としている．2-gram モデル構築に用いるデータには，A，B それぞれの動画像の約 25 分間 (50000 フレーム) を用いている．また，2-gram モデルの構築および適切度 (Relevance) の計算は，ケンブリッジ大学の Toolkit[15] を用いている．実験結果を表 3 に示す．

表 3 において，Boundary detected は検出されたシーン境界の数，Correct boundary は検出されたシーン境界のうち正しいシーン境界の数，Precision はシーン境界の適合率を示している．表 3 において，適合率はまずまず高い値を示しており，シーン検出アルゴリズムが有用であると考えられる．キャッシュモデルを用いた場合の検出数が多いのは，2.2.1 で定義した 2 種類の重み係数を用いたキャッシュモデルの値それぞれに対してシーン境界を検出したためである．

キャッシュモデルのみを用いた場合の適合率と，動画像区間を導入してマージングを行った場合の適合率にはそれほど大きな違いはない．この結果から，動画像区間を導入することで精度を落とすことなくより広い範囲の意味的構造を発見できると考えられる．

画面の全体から情報を抽出した場合と外側の情報のみを抽出した場合の結果には，ほとんど変化が見られなかった．検出したシーン境界点もほぼ一致しており，明確な違いは得られなかった．原因としては，まず，外側として分割した領域が大きすぎたことが考えられる．今回の実験で用いた領域は全体の 4 分の 3 を占めており，背景だけでなく人物などの情報も含んだ可能性が考えられる．もうひとつの原因として，登場人物の位置関係が考えられる．例えば，ニュースではアナウンサーが画面中央に，背景が画面外側に配置された動画像が用いられるが，野球中継では人物は画面の外寄りに位置することが多い．このように，動画像の種類によっては必ずしも画面外側に背景があらわれるとは限らない．今回用いたアニメーションでは，登場人物が画面中央だけではなく，外側にあらわれるショットも多く見られたため，画面全体の情報を抽出した場合とあまり差が見られなかったと考えられる．

一方，2-gram モデルを用いて検出されたシーン境界を実際に見ると，切り出されたシーンの長さが短いために，誤検出となる部分が見られている．本研究では，画面全体と画面外側の情報を用いてシーン検出を行っているが，一方で検出されたシーンが他方で検出されたシーンに含まれる場合がある．このような場合は，マージングを行なって小さな意味のシーンをより大きなシーンにまとめれば，誤検出をおさえることができると考えられる．そこで，画面全体の情報から得られたシーン境界と画面外側の情報から得られたシーン境界とのマージングを行った．実験結果を表 4 に示す．

表4　シーン境界のマージング結果

	Movie A	Movie B
Boundary detected	18	15
Correct boundary	15	12
Precision (%)	83.3	80.0

表 4 の結果を見ると，どちらも高い適合率が得られており，誤検出のシーンが多く取り除かれ，シーンの検出精度が向上していることがわかる．

4. 部分的な動画像からの全体動画像の特定

4.1 基本概念

実験に用いた動画像データは，全体で 2 時間ほどあるオリジナル動画像の一部分を切り出したものである．したがって，オリジナル動画像全体の 2-gram モデルを構築すれば，部分データの 2-gram データはすべて 2-gram モデルに適合する．これを応用すれば，動画像データそれぞれに対して N-gram モデルを構築しておいて，ダイジェストのような部分的な動画像を入力データとし，適合率の高い動画像を検索すれば，部分的な動画像がどのオリジナル動画像に属しているかを特定できると考えられる．本章では，N-gram モデルの応用として，部分的な動画像からオリジナル動画像の特定する方法について述べる．

4.2 実験と評価

4.1 で述べた考え方を用いて，動画像を特定する実験を行った．動画像形式は，第 3 章で述べたものと同じである．オリジナル動画像としては，約 2 時間の 3 本のアニメーション映画 A，B，C の約 25 分間（50000 フレーム）を用い，部分動画像として，A，B，C それぞれの動画像から約 1 分（2000 フレーム）の動画像 a，b，c を，約 5 分（10000 フレーム）の動画像 a'，b'，c' を切り出した．

実験は，以下の手順で行った．

(1) 2.2.2 と同様の手法を用いて，オリジナル動画像 A，B，C それぞれに対して 2-gram モデルを構築する．

(2) それぞれの部分動画像において，ショットの色度数ヒストグラム差分列の離散化を行い，離散化データ列中のすべての 2-gram に対し，先に構築したモデルに含まれているかどうかの検証を行う．

(3) 部分動画像の 2-gram 適合率の最も高い動画像を，オリジナル動画像として検出する．

2-gram モデルの構築および 2-gram 適合率の計算には，3.2 と同じくケンブリッジ大学の Toolkit を用いている．実験結果を表 5 に示す．

表 5 において，Correct A，B，C はそれぞれの部分動画像の 2-gram のうちオリジナル動画像の 2-gram モデルに適合した数，P_A，P_B，P_C はそれぞれの部分動画像のオリジナル動画像に対する適合率を示す．

表 5 の結果より，それぞれの部分動画像の適合率は，もととなるオリジナル動画像に対して最も高くなっており，部分動画像からもととなるオリジナル動画像の特定を行うのが十分可能だと考えられる．オリジナル動画像に対する適合率が 100% とならないのは，切り出した部分動画像がショットの開始点から始まっているとは限らず，最初の色度数ヒストグラム差分値が 2-gram モデルと適合しないためである．そのため，部分動画像に含まれるショット数が非常に少ない場合，例えば 5 ショット程度の場合は，各オリジナル動画像に対する適合度に明確な差が見られず，オリジナル動画像の特定は難しいと考えられる．

表5　部分動画像と全体動画像の検証

	a	b	c	a'	b'	c'
2-gram	14	13	17	65	64	63
Correct A	12	3	10	63	33	31
Correct B	1	11	8	29	62	30
Correct C	7	5	15	34	51	61
P_A(%)	85.7	23.1	58.8	96.9	51.6	49.2
P_B(%)	7.1	84.6	47.1	44.6	96.9	47.6
P_C(%)	50.0	38.5	88.2	52.3	79.7	96.8

5. まとめと今後の課題

本稿では，動画像データに含まれる意味的情報を発見して，インデックス付けを支援することを目的とした，動画像の意味的構造と発見アルゴリズムについて述べた．動画像の意味的構造の発見では，ショットに含まれる情報を用いて意味的なまとまりをあらわす動画像区間を再構築した．本アルゴリズムでは，音声言語の生成確率モデルであるキャッシュモデルを用いて，ショット間の色度数ヒストグラム差分列から 2 種類の動画像区間を定義している．さらに，キャッシュモデルと同じく，音声言語の生成確率モデルである N-gram モデルを用いて，生成確率の変化からシーンの境界点を検出している．

最後に，評価実験によって，実際に動画像における意味的構造を発見し，意味的構造の発見アルゴリズムの有用性を検証した．実験結果から，人手を必要とせずに動画像から意味的情報をもつまとまった動画像区間を発見できたことがわかった．また，確率モデルの応用例として，部分的な動画像からもととなるオリジナル動画像を特定する方法について述べ，有用性を検証した．

今後の課題としては，一般的な動画像の N-gram モデルの構築があげられる．今回の実験で用いた N-gram モデルは，対象となる動画像を含む，より長い動画像を用いて構築している．すなわち，対象動画像に特化

した N-gram モデルとなっている．音声認識の分野の言語モデルは，複数の元データからの大量のコーパスをもとに，統計的に情報源をモデル化してシステムを構築している．動画像の色情報の変化は，その動画像を作った人や，映画やドキュメンタリーなど動画像のタイプによって異なるため，一般的なモデルを構築する場合には，ある程度，動画像を分類した上で構築する必要があると考えられる．

また，動画像からショットの先頭フレーム画像の色度数ヒストグラムを抽出し，ショット間の色度数ヒストグラムの差分を離散化したものを確率モデルへの入力データとして用いたが，色度数ヒストグラムを別の系列データに変換したり，色情報ではなく輝度情報を用いるなど，動画像情報から入力データへの符号化の手法について考えることも検討課題として残されている．

参 考 文 献

1) 長坂晃朗, 田中譲, “カラービデオ映像における自動索引付け法と物体探索法,” 情報処理学会論文誌, Vol.33, No.4, pp.543–550 (1992).
2) K. Hong, J. Tanahashi, M. Kusaba and S. Sugita, “A Motion Picture Archiving Technique and Its Apprication in an Ethnology Museum,” Proc. of 3rd Intl. Conf. on Database and Expert Systems Applications (DEXA92), pp.209–214 (1992).
3) D. H. Watcher, T. Kanade, M. A. Smith and S. M. Stevens, “Intelligent Access to Digital Video,” Informedia Project, IEEE Computer, Vol.29, No.5, pp.46–52 (1996).
4) G. A. Hauptmann and D. Lee, “Topic Labeling of Broadcast News Stories in the Informedia Digital Video Library,” Proc. of 3rd ACM Digital Libraries, pp.287–288 (1998).
5) 有木康雄, “DCT 特徴のクラスタリングに基づくニュース映像のカット検出と記事切り出し,” 電子情報通信学会論文誌, Vol.J80-D-II, No.9, pp.2421–2427 (1997).
6) 中村哲, 北研二, 永田昌明, “音声言語の確率モデル,” 人工知能学会誌, Vol.10, No.2, pp.17–24 (1995).
7) D. Beeferman, A. Berger and J. Lafferty, “Statistical Models for Text Segmentation,” Machine Learning, Vol.34, No.1–3, pp.177–210 (1999).
8) 是津耕司, 上原邦昭, 田中克己, “映像の意味的構造の発見,” 情報処理学会論文誌, Vol.41, No.1 (印刷中).
9) R. Lienhar, “Automatic Text Recognition for Video Indexing,” Proc. of 4th ACM Multimedia, pp.11–20 (1996).
10) Y. Ariki, E. Iwanari and Y. Motegi, “Detection and Description of TV News Article,” Proc. of the 47th FID, pp.198–202 (1994).
11) K. Tonomura, “Video Handring based on Structured Information for Hypermedia Systems,” Proc. of IEEE Computer Graphics & Applications, pp.67–74 (1991).
12) G. Thomas, A. Smith and G. Davenport, “The Stratification System: A Design Environment for Random Access Video,” Proc. of Workshop on Networking and Operating System Support for Digital Audio and Video, pp.250-261 (1992).
13) G. Davenport, G. Thomas, A. Smith and N. Pincever, “Cinematic Primitives for Multimedia,” Proc. of IEEE Computer Graphics & Applications, pp.67–74 (1991).
14) 谷澤和昭, “視覚的内容に基づく動画像のクラスタリングとシーン検出,” 1997 年度神戸大学工学部情報知能工学科卒業論文 (1998).
15) http://svr-www.eng.cam.ac.uk/prc14/toolkit.html