

## 映像の意味的構造の発見

是 津 耕 司<sup>†</sup>, 上 原 邦 昭<sup>††</sup> 田 中 克 己<sup>†††</sup>

ある出来事や情景といった意味の情報に基づいて映像データにアクセスするためには、内容記述によるインデックス付けが必要である。内容記述によるインデックス付けで最も重要なことは、映像データの中から意味的信息を表している映像区間をいかにして特定するかである。本論文で提案する意味的構造とその発見メカニズムは、映像区間を意味的なまとまりを持つ連続したショット列として定義し、これに基づいてショットに分割された映像データから映像区間を見つけ出し、その内容記述を生成するものである。本手法の特徴は、従来の記述モデルのように、映像区間を記述者の手作業によって事前定義するのではなく、発見メカニズムによって映像区間とその内容記述を動的に定義できることである。さらに、提案手法の応用として、映像の意味的構造に基づく映像 skimming を用いた映像データのブラウジング手法について述べる。

### Discovering Semantic Structures of Video Data

KOJI ZETTSU,<sup>†</sup> KUNIAKI UEHARA<sup>††</sup> and KATSUMI TANAKA<sup>†††</sup>

Indexing video data based on contents annotation can fully explore semantic information of video data. However, the most difficult and time-consuming process in annotation-based indexing is to identify appropriate video intervals for various semantic information manually. Thus, discovering video intervals from video data automatically will be helpful for the indexing work. For this purpose, we propose a mechanism for discovering “semantic structures” of video data. This mechanism is to discover video intervals as consecutive sequences of video shots, each of which represents an action or a situation, and also generate annotations of discovered video intervals from annotations of shots. The most significant characteristic of our approach is an author or librarian of video databases can continue his/her indexing work without identifying video intervals previously. An application for quick video data browsing based on semantic structures is also introduced.

#### 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・加工できるようになってきた。それにとともない、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

いう要求が高まってきている。

映像データは、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。映像データの特徴は、画像や音声といった物理的な情報から、それによって表現される出来事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、映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。これまでに、映像データの画像情報や、文字認識あるいは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が、デジタルライブラリーシステム<sup>17),18)</sup>や News On Demand システム<sup>19),20)</sup>などで行われている。

一方、映像データを意味的信息、すなわち映像によって表現される出来事や情景などに基づいてインデックス付けする場合には、意味的信息を表現している時間区間の映像(映像区間)を特定し、その内容を記述することが必要である。このような映像区間の特定には

<sup>†</sup> 通信・放送機構神戸リサーチセンター  
Kobe Research Center, Telecommunications Advancement Organization of Japan  
現在、日本 IBM e-ビジネスソリューション技術  
Presently with e-business Solution Technology, IBM Japan

<sup>††</sup> 神戸大学都市安全研究センター  
Research Center for Urban Safety and Security, Kobe University

<sup>†††</sup> 神戸大学大学院自然科学研究科  
Graduate School of Science and Technology, Kobe University

知識や経験が必要であるため、これまでインデックス付けを行う記者が手作業で映像区間の定義を行っていた。しかし、あらゆる意味の情報に対する映像区間を記者の手作業によって定義するのは大変な作業であり、このことが意味の情報に基づく映像データのインデックス付けを難しいものにしてきた。

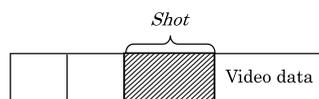
こうした問題を解決するため、本研究では、映像データの中から何らかの意味的なまとまりを表している映像区間を発見し、その内容記述を生成するメカニズムの開発を行っている。本メカニズムは、意味の情報に基づいてシーン検索などを行う映像データベースにおいて、インデックス付け作業の支援システムとして利用することができる。また、映像の要約システムを開発するうえでの記述支援アプリケーションとしても利用可能である。このような記述支援アプリケーションでは、記者は提案手法によって自動生成されたインデックスを使い、必要であればそれらを修正しながらインデックス付け作業を進めていくことができる。このため、従来はすべて手作業で行わなければならないインデックス付け作業の負担を、大幅に軽減することができると考えられる。

## 2. 背景

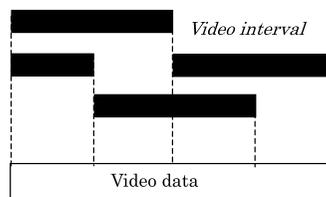
映像データは時間的な連続性を持った情報メディアであり、どの時間区間の映像を見たかによってその内容が変化する。そのため、映像データのインデックス付けでは、映像データを適切な時間区間に区切り、それぞれに対してインデックスを付けることが必要である。インデックス付け手法は、映像データの区切り方によって、structured approach<sup>3)</sup>と stratified approach<sup>1),2)</sup>に分類される。

Structured approach では、映像データを、ショットと呼ばれる時間的な重なりのない連続した時間区間に分割する(図1(a))。ショットへの分割は、フレーム画像間の色の变化など、主に映像データの物理的特徴量に基づいて行われ、各ショットはこれらの特徴量を用いてインデックス付けされる。Structured approach の特徴は、インデックス付け作業を、ショット検出アルゴリズムなどを用いてほぼ完全に自動化できることである。

これに対し、stratified approach では、ある出来事や情景など、意味的なまとまりを表す時間区間の映像ごとに映像区間を定義し、それぞれの映像区間の内容をキーワードなどを使って記述したものをインデックスとして付与する(図1(b))。Stratified approach では、映像区間が互いに時間的な重なりを持つことを許



(a) Structured approach



(b) Stratified approach

図1 時間区間に基づく映像データの管理

Fig. 1 Video data management based on temporal intervals.

しているために、意味的なまとまりに応じて自由にインデックス付けができる。しかし、これらの映像区間の定義は、今日の画像・音声処理技術などを用いても自動化することは難しく、記者によってすべて手作業で行われているのが現状である。

## 3. 映像の意味的構造の発見

### 3.1 基本概念

映像データは、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繋ぎあわせて構成されている。したがって、ショットの内容を比較し、ある出来事や情景を表すように構成されたショット列を見つけ出すことができれば、意味的信息を表す映像区間を定義できる。このような考え方に基づいて、structured approach と stratified approach を統合するアプローチを考案した。本アプローチでは、まず、structured approach に従って映像データをショットに分割する。さらに、各々のショットの内容に基づいて意味的なまとまりを表している連続したショット列を発見し、それらのショット列を stratified approach における映像区間として定義する。これにより、これまで記者の手作業に頼ってきた意味的信息の特定が発見メカニズムによって行えるようになり、映像区間の定義にかかる負担や記者による定義のばらつきを抑えることができる。

図2に、我々のアプローチの概要を示す。図2は、意味的なまとまりを表す映像区間とショット列の対応付けを定義し、ショットからこれらの映像区間を再構成するメカニズムと、発見された映像区間の内容記述を生成するためのメカニズムを示している。

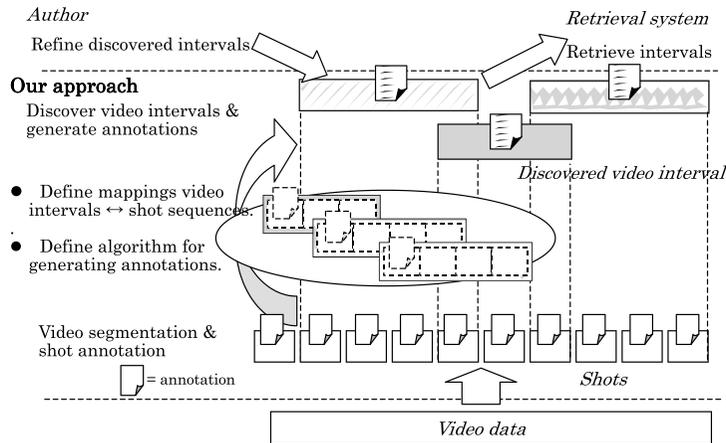


図 2 提案手法の概要  
Fig. 2 Basic concepts of our approach.

3.2 映像の意味的構造

映像の意味的構造とは、ショットの内容に基づいて発見することができる、意味的なまとまりを表す映像区間のことである。意味的構造の中で定義される映像区間は、以下の3種類に分類される。これらの分類は、映像編集におけるショット繋ぎの経験的知識に基づいており、多くの映像において、あるまとまった出来事や情景を表すためにこれらのパターンが用いられている<sup>13)</sup>。

**Unchanged:** 登場人物や背景などが映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。内容記述は、映像区間の先頭の内容によって特徴付けられる。

**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。内容記述は、映像区間の中で顕著な変化を見せる内容によって特徴付けられる。

**Multiplexing:** 個々の登場人物や背景などがショットごとに交互に現れることによって、各々の繰返しの中で表現されている出来事や情景を互に関連付ける。内容記述は、各々の繰返しで表現される内容の集合として表される。

映像の意味的構造の発見とは、映像データの中からこれら3種類の映像区間を発見することである。意味的構造の発見は以下の手順で行われる。

(1) ショットの内容記述

映像データをショットに分割し、それぞれのショットの内容記述を行う。ショット分割には、フレーム画像の色分散情報に基づくショット検出アルゴリズム<sup>10),11)</sup>を用いている。また、シヨッ

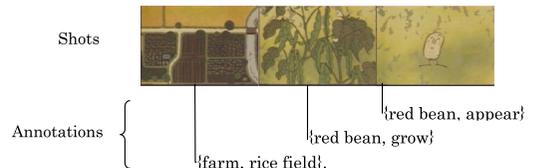


図 3 ショットの内容記述  
Fig. 3 Shot annotations.

トの内容記述は、記述者による記述の差異をできるだけ抑えるために、以下の指針に基づいて行われる。

- 内容記述は、各ショットに表れる登場人物、背景、およびアクションを記述するキーワードの集合として表す。
  - 登場人物の記述では、映像の内容の中心人物の名前を書く。
  - 背景の記述では、映像の背景に登場する物や景色を書く。
  - アクションの記述では、登場人物が行っている行為を表すキーワードを書く。
- 同じ登場人物、背景、アクションに対しては、同じキーワードを使う。

ショットの内容記述例を図3に示す。図3において、1番目のショットには、背景に映し出されている「畑 ( farm )」と「田んぼ ( rice field )」が記述されている。また、2番目のショットには、登場人物の「あずき ( red bean )」とその行為 ( grow ) が記述されている。さらに3番目のショットにも、同様に登場人物の「あずき」とその行為 ( appear ) が記述されている。

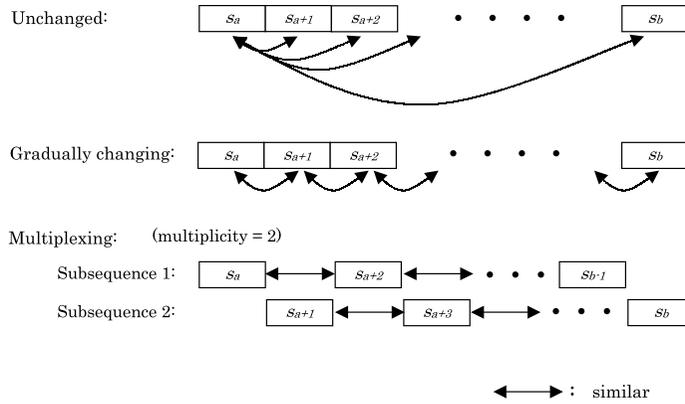


図 4 類似したショットの出現パターンによる映像区間の定義  
Fig. 4 Mappings between video intervals and similar shots.

(2) 映像区間の発見

それぞれの種類の映像区間に対して、映像区間とショット列とのマッピングを定義する。映像区間の発見メカニズムは、ショットに分割された映像データの中から、いずれかのマッピングを満たすショット列を見つけ出し、対応する種類の映像区間として定義している。

(3) 発見された映像区間の内容記述の生成

発見された映像区間の内容記述を、その中に含まれるショットの内容記述から生成する。映像区間の内容記述は、ショットの内容記述に含まれるキーワードの中から、この映像区間の内容を最もよく特徴付けるものを選び出して作成する。そのため、それぞれの種類の映像区間に対して、キーワードの選択基準が定義されている。

3.3 映像区間の発見

映像の意味的構造で定義される 3 種類の映像区間は、ショット間の内容の変化を調べれば発見することができる。それぞれの種類の映像区間は、内容の類似したショットの出現パターンとして以下のように定義している。

- もし  $\forall s \in I, similarity(s, start(I)) > \theta$  であれば、映像区間  $I$  は *unchanged* である。
- もし  $\forall s_i, s_{i+1} \in I, similarity(s_i, s_{i+1}) > \theta$  であれば、映像区間  $I$  は *gradually changing* である。
- もし  $\forall s_i, s_{i+m} \in I, similarity(s_i, s_{i+m}) > \theta$  であれば、映像区間  $I$  は *multiplexing* である。ここで、 $m$  は多重度であり、 $m$  ショットごとに類似したショットが現れることを意味している。多重度に基づいて言い換えると、映像区間  $I[a, b]$  は  $m$  個の subsequence  $\{I_i | 1 \leq i \leq m\}$  が交互に組み合わせられて構成されていることになる。個々

の subsequence は、 $I'_i = [s_{a+(i-1)}, s_{a+(i-1)+m}, s_{a+(i-1)+2m}, \dots, s_{b-m+i}]$  と定義される。

ここで、 $similarity()$  は 2 つのショット間の類似度を計算する類似度関数である。 $\theta$  は計算された類似度に基づいてショット間の類似性を判断するための類似度閾値である。もし 2 つのショット  $s_i, s_j$  の類似度が  $similarity(s_i, s_j) > \theta$  であれば、ショット  $s_i$  と  $s_j$  は類似していると見なされる。これらの定義を図 4 に示す。

ショットの内容記述に基づいて 3 種類の映像区間を発見する様子を図 5 に示す。映像区間の発見は、以下の手順で行われる。

- (1) あらゆるショットの組合せに対して類似度を計算する。類似度関数  $similarity()$  は、ベクトル空間法に基づき、キーワードベクトル形式で表したショットの内容記述間のコサイン相関値を計算する<sup>7)</sup>。キーワードベクトルとは、特定のキーワードが内容記述の中に存在しているかどうかをベクトル形式で表記したもので、各要素に対応付けられたキーワードが存在すれば 1 を、存在しなければ 0 の値をとる。図 5 では、ショット間の類似度が行列形式で表示されている。類似度行列の各要素は、行および列成分に対応する 2 つのショット間の類似度を表している。
- (2) 類似度閾値  $\theta$  を与えて、ショットの類似性を判断する。図 5 では、類似度行列の中で類似度が類似度閾値 (= コサイン相関値 0.2) より大きい要素が灰色で色分けされている。すなわち、この要素の行列成分に対応する 2 つのショットが類似していることを示している。
- (3) 3 種類の映像区間に対応する類似ショットの出現パターンを探索し、いずれかのパターンを満

Similarity between each shot (threshold = 0.2)

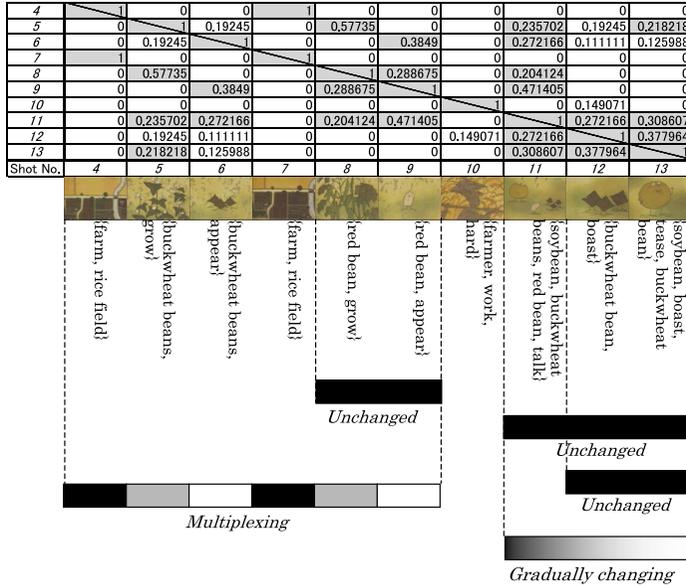


図5 意味的構造の発見メカニズム

Fig. 5 A mechanism for discovering video data.

たすショット列が見つければ、対応する種類の映像区間として定義する。類似ショットの出現パターンの探索は、直感的に、図5に示された類似度行列上で対角成分を順に開始点としながら、色分けされた要素の配列パターンを行方向に探していくことと同じである。3つの配列パターンのうちいずれかが見つければ、この要素配列の行列成分に対応するショット列がパターンに対応する種類の映像区間として定義される。

図5において、発見された映像区間の中で相互に時間的な重なりを持ったものがあるのは、映像区間の開始位置やその長さによって意味的情報が異なるという、意味的情報の時間依存性を表している。

3.4 映像区間の内容記述の生成

発見された映像区間の内容記述は、映像区間に含まれるショットの内容記述に使われているキーワードの中から、その映像区間の内容を特徴付けるキーワードを選択し生成される。これは、ショットの内容記述に含まれるキーワードをランク付けし、高いランクに位置するキーワードを選択することに相当している。キーワードの選択基準は、それぞれの映像区間の種類ごとに以下のように定義している。なお、 $I$ は発見された映像区間を示し、 $v(I)$ は $I$ を構成するショットの内容記述に含まれるキーワードを要素として持つキーワードベクトルを表している。 $v(I)$ の各要素は、選択基準によって計算される各キーワードの重み付けを表し

ている。キーワードのランク付けは、この重み付けに基づいて行われる。

**Unchanged:**  $v(I)$ は、 $I$ に含まれる各ショットの内容記述を表すキーワードベクトルに対し、 $I$ の先頭ショットと各ショットとの類似度によって重み付けした、重み付け平均ベクトルである。したがって、 $v(I)$ では、 $I$ の先頭ショットの内容が最も強調される。

$$v(I) = \frac{\sum_{i=start(I)}^{end(I)} similarity(s_i, s_{start(I)})v(s_i)}{\sum_{i=start(I)}^{end(I)} similarity(s_i, s_{start(I)})} \tag{1}$$

$start(I)$  および  $end(I)$  は、それぞれ  $I$  の先頭ショットおよび最終ショットを示す。

**Gradually changing:**  $v(I)$ は、以下の行列  $M$  の最大固有値に対応する固有ベクトルとして定義される。

$$M = \sum_{k=1}^n \delta_k \delta_k^T \tag{2}$$

$\delta_k$  は、ショット  $s_k$  のキーワードベクトル  $v(s_k)$  と  $I$ に含まれるすべてのショットのキーワードベクトルの平均  $\bar{v}$  の差分ベクトルである。

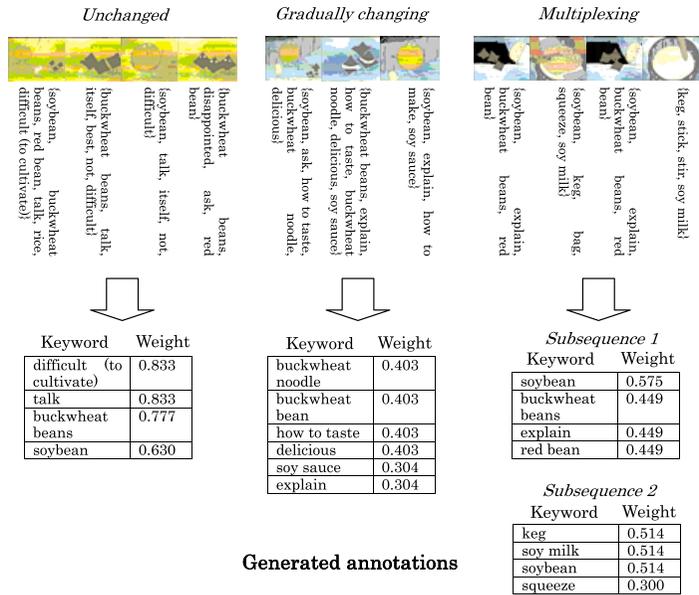


図 6 発見された映像区間の内容記述の生成  
Fig. 6 Generating annotations of discovered video intervals.

$$\delta_k = v_i - \bar{v}$$

$$(start(I) \leq i \leq end(I), 1 \leq k \leq n)$$

$$\bar{v} = \frac{\sum_{i=start(I)}^{end(I)} v(s_i)}{n}$$

$v(I)$  の各要素は、 $I$  全体の平均からの変化の大きさに応じて重み付けられるため、 $v(I)$  は、直感的に、 $I$  の中における変化を特徴付けている。

**Multiplexing:**  $I$  を構成する各 subsequence ごとに、*gradually changing* と同じ方法でキーワードベクトルを作成し、キーワードのランク付けと選択を行う。

図 6 に、発見された映像区間に対する内容記述の生成の様子を示す。図 6 には、選択基準に従って選択されたキーワードがその重み付けとともに示されている。*Unchanged* タイプの映像区間の内容記述は、先頭ショットに現れるキーワードが強調されている。*Gradually changing* タイプの映像区間の内容記述は、頻繁に出現・消滅を繰り返すキーワード、あるいは出現・消滅の状態のいずれも長いキーワードが特に強調され、映像区間全体の内容の特徴付けている。*Multiplexing* タイプの映像区間の内容記述では、個々の subsequence ごとに選択されたキーワードが各々の内容を表している。

#### 4. 意味的構造発見の評価

映像の意味的構造の発見メカニズムを評価するために、被験者として大学からのボランティア 10 名に参

加してもらい、実際の映像データを使って評価実験を行った。実験には、約 10 分間のアニメーション映像を用いた。この映像データをフレーム画像色分散の  $\chi^2$  検定に基づくカット検出プログラム<sup>11)</sup>により 79 ショットに分割し、被験者が手作業で各ショットの内容記述を行った。

##### 4.1 発見メカニズムの評価

###### 4.1.1 映像区間の種類に関する評価

意味的構造で定義した 3 種類の映像区間 (*unchanged*, *gradually changing*, *multiplexing*) が十分に汎用的であるかどうかを評価するため、発見メカニズムによって検出された映像区間が映像データ全体に占める割合を調べた。図 7 に結果を示す。“Threshold” は類似度閾値を、“Coverage” は発見された映像区間が映像全体に占める割合 (ショット数に基づく) を示している。図 7 において、発見メカニズムによって発見された映像区間が映像全体に占める割合 (coverage) は、53~87% と十分に高い値を示している。これは、我々が定義した 3 種類の映像区間が十分に汎用的であり、映像データの大部分をカバーできることを示している。

###### 4.1.2 映像区間の発見に関する評価

映像区間をどの程度正確に発見できるかを評価するため、3 種類の映像区間について、被験者が手作業で検出した映像区間と発見メカニズムが検出した映像区間とを比較し、被験者が検出した映像区間の中で発見メカニズムが検出でなかった映像区間数の割合 (検出エラー率) を調べた。被験者には、意味的構造の定義

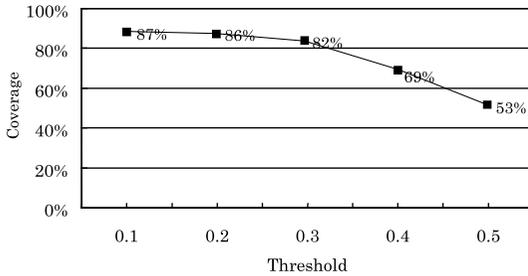


図 7 発見された映像区間の coverage

Fig. 7 Coverage of discovered video intervals.

に基づき次の指針に従って 3 種類の映像区間の検出を行ってもらった。すなわち, *unchanged* はすべてのショットが類似している映像区間, *gradually changing* は隣接するショットが類似している状態が続く映像区間, *multiplexing* は類似したショットが交互に繰り返し現れる映像区間を見つけ出す。なお, 被験者と発見メカニズムの間で映像区間の検出条件を揃えるため, ショット間の類似性判断に 2 つの基準を設けた。『ゆるい類似基準』では, 被験者は登場人物, 背景, アクションを表すキーワードのうち 1 つでも同じものがあれば 2 つのショットは類似していると思わずのに対し, 発見メカニズムは類似度閾値を 0.2~0.3 に設定する。一方『厳しい類似基準』では, 被験者は登場人物, 背景, アクションを表すキーワードのうち 8 割程度同じものがあれば 2 つのショットは類似していると思わずのに対し, 発見メカニズムは類似度閾値を 0.5 に設定する。

各被験者が作成したショットの内容記述に基づいて実験を行った結果, ゆるい類似基準における検出エラー率は被験者全体で 29%, および厳しい類似基準における検出エラー率は同じく 49%であった。この結果から, 提案した発見メカニズムは, 内容記述支援を目的として映像区間を発見するためにまずまずの性能を示していることが分かる。

#### 4.1.3 発見された映像区間の分類に関する評価

発見された映像区間がどの程度正確に分類されるかを評価するため, 検出された映像区間を各種別の映像区間ごとに比較し, 被験者が分類した映像区間と発見メカニズムが分類した映像区間のうちで, 分類が異なる割合(分類エラー率)を調べた。なお, 分類には, 先ほどの実験と同じ 2 つの類似性判断基準を用いた。

実験結果を表 1 に示す。*Gradually changing* に関しては分類エラー率が大きくなっている。これは, 被験者のコメントにもあるように, 人間が必ずしも内容の連続性にのみ着目して意味的まとまりを検出するの

表 1 発見された映像区間の分類

Table 1 Classification of discovered video intervals.

映像区間の種類	分類エラー率	
	ゆるい類似基準	厳しい類似基準
Unchanged	30%	55%
Gradually changing	57%	88%
Multiplexing	29%	65%

ではなく, 内容が途切れる箇所を手がかりとして, それまでの映像区間を *gradually changing* として検出していることによる。これに対し, 発見メカニズムは, 類似した内容の連続性に関するパターン(3.3 節)に基づいて映像区間の発見および分類を行っている。このため, 内容を次々と連続させてストーリー展開を表現するような *gradually changing* では, 両者で分類が大きく違っていると考えられる。

#### 4.2 内容記述生成の評価

##### 4.2.1 生成された内容記述に関する評価

発見された映像区間の内容記述の生成アルゴリズムを評価するため, 生成された内容記述がどの程度正しいかを調べた。データ検索においてインデックスの正確さを評価するために広く用いられている適合率を以下のように修正し, 生成された内容記述の正確さの判断基準として用いた。

$$\text{適合率} = \frac{\left( \frac{\text{生成された内容記述が正しい映像区間の数}}{\text{発見された映像区間の数}} \right)}{\text{発見された映像区間の数}} \quad (3)$$

被験者が作成したショットの内容記述に基づいて, 生成アルゴリズムが生成した内容記述の適合率を算出した結果を表 2 に示す。“Precision ratio” は, 生成された内容記述の適合率を映像区間の種類別に示している。なお, 適合率の計算の際, 生成された内容記述が正しいかどうかは, 対応する映像区間の映像を各被験者が実際に見て判断した。表 2 において, 適合率は十分に高い値を示しており, 生成された内容記述の多くが正しく内容を記述していることが分かる。一方, 生成された内容記述が不適切であると判断された映像区間の中には, 被験者が映像区間の中で印象に残った特定の内容に基づいて内容記述を行おうとしている場合が多いことが分かった。たとえば, 登場人物らが会話をしている映像区間に対し, 生成アルゴリズムは登場人物や話すというアクションに対するキーワードを優先的に選択するのに対し, 被験者は会話中の特定の話題を映像区間全体を代表するキーワードとして選択する。現在の生成アルゴリズムは, 映像区間の種類に応じて特定の位置にあるキーワードの重み付けを大き

表 2 生成された内容記述の適合率

Table 2 Precision ratio of generated annotations.

Video interval	Precision ratio
Unchanged	75%
Gradually changing	85%
Multiplexing	56%

表 3 映像区間と選択基準の組合せによる適合率の変化

Table 3 Precision ratio of discovered intervals against selection criteria.

Selection criteria	Precision ratio		
	Unchanged	Gradually changing	Multiplexing
Unchanged	0.739	0.625	0.105
Gradually changing	0.696	0.875	0.316
Multiplexing	N/A	N/A	0.579

くしているが、今後はさらに変化の特徴に応じて可变的に重み付けを変えるアルゴリズムも必要であると考えられる。

4.2.2 キーワードの選択基準に関する評価

内容記述の生成の際に適用されるキーワードの選択基準が、それぞれの種類の映像区間に対し最適に定義されているかについて評価を行った。ある被験者の内容記述に基づいて発見メカニズムにより発見された各映像区間に対し、その映像区間の種類に対応した選択基準とそれ以外の種類に対する選択基準を適用し、それぞれの選択基準に基づいて生成された内容記述の適合率を計算した。表 3 に結果を示す。“Selection criteria”は選択基準の種類を、“Precision ratio”はそれぞれの種類の映像区間に対し“Selection criteria”に示された選択基準を使って生成された内容記述の適合率を示している(なお、multiplexingの選択基準は映像区間が2つ以上のsubsequenceから構成されていることを前提としているため、subsequenceの定義のないunchangedやgradually changingの映像区間に対しては“N/A”と記されている)。表3では、いずれの種類の映像区間においても、発見された映像区間と同じ種類の選択基準を用いたときに適合率が最も高くなっている(表3の対角成分)。この結果から、選択基準はそれぞれの種類の映像区間に対して最適に定義されていることが分かる。

5. 意味的構造に基づく映像 skimming

本章では、意味的構造の発見メカニズムの応用として、意味的構造に基づく映像 skimming による映像データのブラウジング手法について述べる。

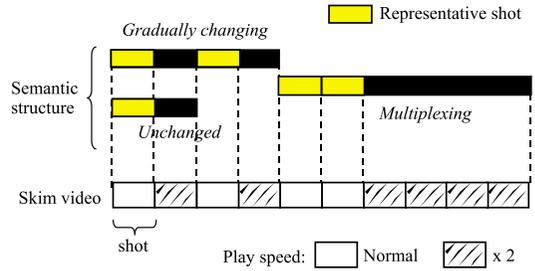


図 8 意味的構造に基づく映像 skimming

Fig. 8 Video skimming based on semantic structures.

5.1 基本的な考え方

映像データの skimming とは、映像データの中から重要な部分だけを抜き出した短縮版 (skim 映像) を作成し、映像データ全体の内容を短い時間でブラウズできるようにすることである。意味的構造に基づく映像 skimming では、意味的構造の発見によって定義された映像区間、すなわちある出来事や情景といった意味的まとまりごとに、代表的なショットを抜き出して再生し、それ以外の重要でない部分を早送りすることにより映像を圧縮する。そのため、様々な出来事や情景の重要な部分とそれらの展開を追いながら、映像全体の内容を短時間でブラウズすることができる。

意味的構造に基づく映像 skimming を図 8 に示す。図 8 で、“Semantic structure”は意味的構造を、“Representative shot”は発見された各映像区間の内容を代表するショットを示している。また、“Skim video”は意味的構造に基づいて作成された短縮版の skim 映像を表しており、“Normal”で示されたショットは通常で再生され、“x 2”で示されたショットは早送り再生される。

意味的構造の発見メカニズムでは、図 7 で示されたように、ショット間の類似度閾値を変えることにより、意味的なまとまりを表す 3 種類の映像区間によってカバーされる映像データの範囲が動的に変化する。たとえば、閾値が低い場合は、発見された映像区間によって映像データのほぼ全体がカバーされるが、逆に閾値を高くすれば、より特徴的な部分のみがカバーされる。したがって、意味的構造に基づく映像 skimming では、類似度閾値を変化させることにより、映像データの圧縮率を動的に調節することができる。ユーザは類似度閾値により圧縮率を変えながら、始めは高い圧縮率で映像全体の内容を大まかに把握し、その後圧縮率を下げて、興味のある部分の内容を詳細に見るといったようなことができる。

5.2 skimming メカニズム

意味的構造に基づく映像 skimming は、以下の手順

で行われる．

- (1) 意味的構造により発見された各々の映像区間の中から，それぞれの種類の映像区間ごとに，以下の方法によって代表ショットを選び出す．

**Unchanged**：映像区間の先頭ショットを選択する．

**Gradually changing**：映像区間の中で，登場人物や背景などの内容が変化する時点のショットを選択する．すなわち，映像区間の先頭ショットを最初の基点として，後続のショットの中で現在の基点ショットに類似していない（類似度が類似度閾値以下の）最初のショットを選択し，これを映像区間の代表ショットに含める．さらにこのショットを新たな基点として，映像区間の終わりに達するまでこの操作を繰り返す．

**Multiplexing**：映像区間を構成する subsequence ごとに，*gradually changing* と同じ方法で代表ショットを選び出す．

- (2) 相互に重なりを持った映像区間をマージする．

- 2つの映像区間  $I_A$  と  $I_B$  が overlap している場合，すなわち  $start(I_A) < start(I_B)$  かつ  $end(I_A) < end(I_B)$  の場合，これら2つの映像区間をマージして新しい映像区間  $I_{AB}$  を作る．ここで， $start(I_{AB}) = start(I_A)$  かつ  $end(I_{AB}) = end(I_B)$  である． $I_{AB}$  の代表ショット  $rep(I_{AB})$  は  $rep(I_{AB}) = \{s_i | s_i \in rep(I_A), s_i < start(I_B)\} \cup rep(I_B)$  となる．

- 映像区間  $I_A$  が映像区間  $I_B$  を含む場合，すなわち  $start(I_A) < start(I_B)$  かつ  $end(I_A) > end(I_B)$  の場合， $I_A$  に  $I_B$  を統合する．この場合，代表ショットは  $I_A$  のものを用いる．

$start(I)$  および  $end(I)$  は映像区間  $I$  の開始ショットおよび終了ショットを示す．

- (3) 各ショットを以下のように再生する．

- (2) のマージ後に残った映像区間の代表ショットは通常どおりに再生する．
- (2) のマージ後に残った映像区間に含まれる代表ショット以外のショットは早送り再生する．
- (2) のマージ後に残った映像区間に含まれないショットは通常どおりに再生する．

### 5.3 映像 skimming の評価

意味的構造に基づく映像 skimming のプロトタイプ

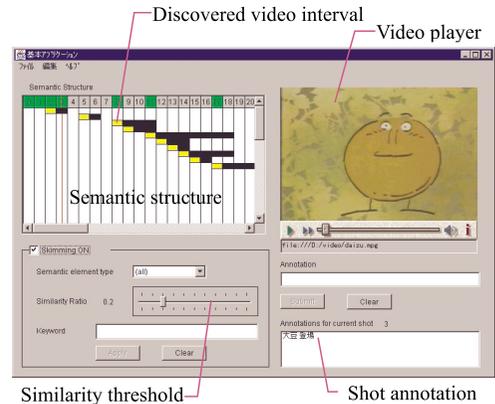


図9 意味的構造に基づく映像 skimming のプロトタイプシステム  
Fig.9 Prototype system of video skimming based on semantic structures.

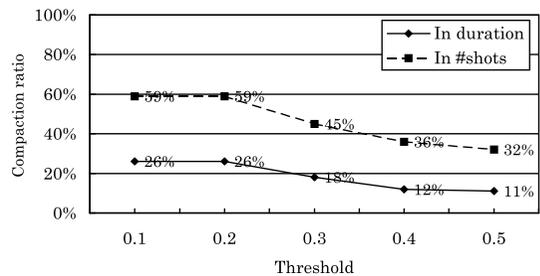


図10 映像 skimming による映像の圧縮率  
Fig.10 Compaction ratio in video skimming.

システムを作成し，実際の映像を使って評価実験を行った．図9にこのプロトタイプシステムを示す．

映像 skimming によってどの程度映像が圧縮されるかを調べた結果を図10に示す．図10では，横軸“Threshold”が類似度閾値を，縦軸“Compaction ratio”が圧縮率を示す．また，“In duration”は再生時間で見た圧縮率の変化を，“In #shots”はショット数で見た圧縮率の変化を示している．ショット数による圧縮率に関しては，十分な圧縮率が得られていると考えられる．再生時間による圧縮率に関しては，ショットの早送りの速度を調節することによってさらに高い圧縮率が得られると考えられる（今回の実験では，通常再生の2倍速で早送り）．

さらに，skimmingによって圧縮された映像の中にどのような情報が残されるかについて調べた．Skim映像に含まれる意味的な情報を定量的に測ることは困難であるため，skim映像を見て理解された内容の変化を調べることにより，skimmingによって残される意味的な情報の変化の特徴を調べた．実験は，被験者に粗い skim 映像（図10で threshold が0.2，圧縮率は59%），詳細な skim 映像（同 threshold が0.5，圧

縮率は 32%)、および圧縮されていない映像を見てもらい、各々の映像の内容について分かったことを時系列に箇条書きしてもらった。なお、skim 映像の早送り速度による影響を避けるため、被験者には、早送りされる映像は前後のつながりを見る程度で内容まで詳細に見なくてもよいとガイドした。

圧縮されていない映像と粗い skim 映像とを比較した結果、粗い skim 映像で分かった項目には断片的でお互いに関連性のないものが多いことが分かった。たとえば、個々の場面での出来事や場面の転換点の内容などである。これは、skimming によって、映像区間内で内容が大きく変化するショットや、どの映像区間にも属さずそれだけで独立した内容を表しているショットが残されたからだと考えられる。一方、粗い skim 映像と詳細な skim 映像とを比較した結果、skimming が詳細になると理解される項目が増えるばかりでなく、これまでに分かった項目の要約も起こりうることが分かった。たとえば、粗い skimming では主人公(大豆)が自分を材料に作られる様々な物の説明が個々に理解されていたが、より詳細な skimming ではこれらの説明が行われた状況(自分が役に立つことを示す)が理解され、これらが主人公の自慢話であると要約された。このように、粗い skim では映像全体に含まれる様々な場面が断片的に把握され、skimming を詳細にすると個々の事実を取り囲む状況が理解されて要約が起こり、ストーリーが次第に理解されていく。したがって、skimming では、まず粗い skimming で場面を探し、次に各場面で skimming を次第に詳細にししながら内容を理解していく方法が効果的であると考えられる。

## 6. 関連研究

内容記述に基づくインデックス付け手法としては、これまでに、映像区間の代数関係<sup>4)</sup>や時間関係<sup>5),14)</sup>、あるいは内容記述の階層関係<sup>15)</sup>などに基いて映像データの意味的情報を構造化する方法が提案されてきた。これらの手法では、いずれも映像区間は記者によって事前定義されることを前提としており、それらを合成することによって新しい意味を持つ映像区間を生成することを目的としている。我々のアプローチは、これらの手法に対し、最も基本的な意味的情報に対する映像区間とその内容記述を提供することにより、これらの手法で問題であった記者による映像区間の事前定義の負荷を軽減し、内容記述作業を支援する。

一方、映像区間を事前定義しない記述モデルとして、文献 16) では時刻印付オーサリンググラフを提案している。時刻印付オーサリンググラフでは、グラフを

使って様々な時点における断片的な内容記述とそれらの意味的な関連性を記述し、グラフ探索によって特定の意味的情報に対する映像区間を探し出している。この方法では、断片的な内容記述間の関連付けに時間的な制約がないため、自由に関連付けができる反面、検索された映像区間の中に関係のない映像が含まれてしまうことも多い。これに対し、我々の方法では、映像区間を意味的なまとまりを持つ連続したショット列と定義し、強い時間制約を付けることによって、発見された映像区間と関係のない映像が含まれることを防いでいる。

映像 skimming に関して、Smith ら<sup>6)</sup>は、映像データの中に含まれるテロップやナレーションを手がかりに、重要なキーワードを含む映像部分を抜き出すことで映像の圧縮を行っている。作成された skim 映像は、キーワードに対する断片的な映像を繋ぎ合わせたものであり、テキストや音声による内容の説明が中心であるニュース映像などのブラウジングに適している。これに対し、意味的構造に基づく映像 skimming では、ある出来事や情景といった意味的なまとまりごとに、各々の内容を代表するショットを残し、その他のショットを早送りすることによって映像を圧縮している。そのため、全体の意味的な内容の繋がりが展開を損なわない skim 映像を作ることができ、劇画などストーリー性のある映像のブラウジングに有効である。

## 7. まとめと今後の課題

本論文では、映像データに含まれる意味的情報を発見して内容記述によるインデックス付けを支援することを目指し、映像の意味的構造とその発見メカニズムを提案した。映像の意味的構造の発見では、ショットの内容から意味的なまとまりを表す映像区間を再構成する。発見される映像区間をショットの内容の変化に基づいて 3 種類に分類し、それぞれを類似したショットの出現パターンとして定義した。さらに、発見された映像区間の内容記述をショットの内容記述から生成するために、それぞれの種類の映像区間ごとに、ショットの内容記述に使われているキーワードを映像区間の内容記述として選択するための基準を定義した。最後に、評価実験により意味的構造の発見メカニズムの有効性を検証した。また応用例として、映像データを意味的なまとまりごとに圧縮し、全体の意味的情報を損なうことなく短時間で映像をブラウズする映像 skimming について述べた。

今後の課題として、まずショットの内容記述の自動化があげられる。ショットの中に映されている登場人物、

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認識可能である<sup>8),9)</sup>. 今後は, これらの技術を用いてショットの内容記述を自動化し, 意味的構造の発見メカニズムを完全に自動化することを目指したい.

また, キーワードによる内容記述では, 映像の内容を的確に表現するキーワードの選択が重要になる. 今回は内容記述に使用するキーワードを登場人物, 背景, アクションに関するものに限定し, 意味的構造の発見メカニズムの性能が記述者のキーワードのつけ方に左右されにくくなるよう配慮した. しかし, このような単純なキーワードだけでは表現できる内容に強い制約がかかり, 複雑な映像の内容を的確に表現するのに十分ではない. そこで我々は, より情報量が豊富な映像の信号情報(色情報など)も組み合わせる意味的構造を発見する方法の研究を行っている.

もう1つの課題としては, 映像区間の内容記述の生成における paraphrasing(言い換え)問題がある. 今回提案した手法では, 発見された映像区間の内容記述は, その中に含まれるショットの内容記述から選択したキーワードから生成されている. しかし映像区間の内容記述は, その中に含まれるショットの内容記述の単純な組合せではない場合もある. たとえば, “ナイフとフォークを取る”, “肉を切る”, “肉を口に運ぶ”という一連のショットから構成される映像区間の内容記述は, これらの内容記述の組合せではなく, まったく新しい記述“食事をする”のほうがより適切である. このように, 一連の内容記述をまったく新しい記述で置き換えることを paraphrasing と呼び, 主に AI の分野で研究が行われている<sup>12)</sup>. 言い換えは我々人間の知識や経験に負うところが多く, paraphrasing 問題を完全に解くことは難しいが, 今後この問題についてもさらに研究を進める予定である.

謝辞 本論文は, 通信・放送機構神戸リサーチセンターにおける研究プロジェクト「次世代デジタル映像通信に関する研究」における研究成果をまとめたものである. また, 本研究は一部, 文部省特定領域研究「発見科学」, 文部省重点領域研究(課題番号 08244103), および日本学術振興会未来開拓学術研究推進事業における研究プロジェクト「マルチメディア・コンテンツの高次処理の研究」に拠っている. なお, 映像データの学術利用は愛企画センターのご協力による. ここに記して謝意を表す.

## 参考文献

- 1) Thomas, G., Smith, A. and Davenport, G.: The Stratification System: A Design Environment for Random Access Video, *Proc. Workshop on Networking and Operating System Support for Digital Audio and Video*, pp.250–261 (1992).
- 2) Davenport, G., Thomas, G., Smith, A. and Pincever, N.: Cinematic primitives for multimedia. *Proc. IEEE Computer Graphics & Applications*, pp.67–74 (1991).
- 3) Tonomura, Y.: Video handling based on structured information for hypermedia systems, *Proc. Intl. Conf. on Multimedia Information Systems*, pp.333–344 (1991).
- 4) Weiss, R., Duda, A. and Gifford, D.: Content-Based Access to Algebraic Video, *Proc. IEEE Multimedia*, pp.140–151 (1994).
- 5) Allen, J.F.: Maintaining Knowledge about Temporal Intervals, *Comm. ACM*, Vol.26, pp.832–843 (1983).
- 6) Smith, M.A. and Kanade, T.: Video Skimming for Quick Browsing based on Audio and Image Characterization, Tech-Report CMU-CS-95-186, Carnegie Mellon University (1995).
- 7) Salton, G.: *The SMART Retrieval System – Experiments in Automatic Document Processing*, Prentice-Hall Inc. (1971).
- 8) Lienhar, R.: Automatic Text Recognition for Video Indexing. *Proc. 4th ACM Multimedia*, pp.11–20 (1996).
- 9) Ariki, Y., Iwanari, E. and Motegi, Y.: Detection and Description of TV News Article, *Proc. 47th FID*, pp.198–202 (1994).
- 10) Zhang, H.J., Kankanhalli, A. and Stephen, W.S.: Automatic parsing of full-motion video, *Multimedia Systems*, Vol.1, pp.10–28 (1993).
- 11) 谷澤和昭: 視覚的内容に基づく動画のクラスタリングとシーン検出, 卒業論文, 神戸大学工学部情報知能工学科 (1998).
- 12) Schank, R.: *Dynamic Memory*, Cambridge University Press (1982).
- 13) Arijon, D.: *Grammar of the Film Language*, Silman-James Press (1991).
- 14) Hibino, S. and Rundensteiner, E.: MMVIS: Design and Implementation of a Multimedia Visual Information Seeking Environment, *Proc. 4th ACM Multimedia*, pp.75–86 (1996).
- 15) Oomoto, E. and Tanaka, K.: OVID: Design and Implementation of a Video-Object Database System, *IEEE Trans. on Knowledge and Data Engineering*, Vol.5, No.4, pp.626–643 (1993).

- 16) 是津耕司, 上原邦昭, 田中克己: 時刻印付オーサリンググラフによるビデオ映像のシーン検索, 情報処理学会論文誌, Vol.39, No.4, pp.923-932 (1998).
- 17) Hong, K., Tanahashi, J., Kusaba, M. and Sugita, S.: A Motion Picture Archiving Technique and Its Application in an Ethnology Museum, *Proc. 3rd Intl. Conf. on Database and Expert Systems Applications (DEXA92)*, pp.209-214 (1992).
- 18) Wactlar, D.H., Kanade, T., Smith, M.A., Stevens, S.M.: Intelligent Access to Digital Video: Informedia Project, *IEEE Computer*, Vol.29, No.5, pp.46-52 (1996)
- 19) Hauptmann, G.A. and Lee, D.: Topic Labeling of Broadcast News Stories in the Informedia Digital Video Library, *Proc. 3rd ACM Digital Libraries*, pp.287-288 (1998).
- 20) 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).

(平成 11 年 3 月 4 日受付)

(平成 11 年 11 月 4 日採録)



是津 耕司 (正会員)

1970 年生. 1992 年東京工業大学工学部情報工学科卒業. 同年日本 IBM (株) 入社. 1996~1998 年度, 通信・放送機構神戸リサーチセンター研究員として, デジタル映像通信, マルチメディアデータベースの研究に従事. 現在, 日本 IBM においてコンテンツ管理の研究開発に従事. ACM 会員.



上原 邦昭 (正会員)

1954 年生. 1978 年大阪大学基礎工学部情報工学科卒業. 1983 年同大学院博士後期課程単位取得退学. 大阪大学産業科学研究所助手, 講師, 神戸大学工学部情報知能工学科助教を経て, 同大学都市安全研究センター教授. 情報知能工学科を兼任. 1989 年より 1990 年まで Oregon State University, Visiting Assistant Professor. 1994 年より 1996 年まで神戸大学総合情報処理センター副センター長. 工学博士. 人工知能, 特に機械学習, マルチメディアデータベース, 自然言語によるヒューマンインタフェースの研究に従事. 1990 年度人工知能学会研究奨励賞授賞. 人工知能学会, 電子情報通信学会, 計量国語学会, 日本ソフトウェア科学会, AAAI 各会員.



田中 克己 (正会員)

1974 年京都大学工学部情報工学科卒業. 1976 年同大学院修士課程修了. 1979 年神戸大学教養部助手. 1986 年同大学工学部助教授. 1994 年同大学教授 (情報知能工学科). 1995 年同大学院自然科学研究科専任教授, 現在に至る. 主にデータベースの研究に従事. 1995~1998 年度本会データベースシステム研究会主査. 1996~1998 年度, 通信・放送機構「次世代デジタル映像通信の研究開発」の研究総括責任者, 1996~1998 年度, 文部省科研費重点領域研究「分散発展型データベースシステム技術の研究」の研究代表者. 神戸マルチメディアインターネット協議会会長. 人工知能学会, ACM, IEEE CS 等会員.