情 報 メ デ ィ ア 15－4
（１９９４．　３．　１１）

AV情報構造化技術とその情報要約への応用

大久保 雅且　　中川　透
ohkubo@aether.ntt.jp　　nakagawa@aether.ntt.jp

ＮＴＴヒューマンインタフェース研究所

内容梗概　マルチメディア情報の有効な処理・伝達方式を実現するためには情報の構造化が必要となる。しかし、既存情報の多くは、人間にはその構造を認識できても、内部的な表現としては構造化されていない。本報告では、テレビニュースを対象として、人間の知覚という観点からAV (Audio Visual) 情報の構造化について検討する。まず、情報の送受信者間で暗黙のうちに用いられている情報伝達構造の存在を認知実験によって確かめる。この構造は、提案する情報構造認識モデルに基づいて、進行制御人物提示区間検出と話者交替点検出によって計算することができる。さらに、抽出された構造を用いた要約情報の作成について述べる。

Recognition and Application of Audio Visual Information Structure

Masaaki Ohkubo　and　Toru Nakagawa

NTT Human Interface Laboratories

abstract　This paper proposes audio visual information structure recognition model. This model is inherently a protocol, that is implicitly used between senders and receivers of information such as TV news. Existence of the protocol and validity of the model is demonstrated through cognitive experiments. We show that finding intervals where a flow controlling person such as a news caster is shown, and detecting speaker change points are essential to calculate the structure based on the model. Finally, we describe that effective summary presentation can be made by restructuring the original information.

1．はじめに

画像や音声などの情報のデジタル化技術の発展に伴い、近年マルチメディア情報処理に対する期待が高まっている。マルチメディア情報の有効な処理・伝達方式を実現するためには、情報の適切な部分への分解と、それら部分情報を効果的に組織化するための構造情報（インデクス等）の付加が必要である。このような情報の構造化により、例えば、重要な部分へのアクセスを容易にしたり、特定部分の抽出・省略・置換を行って要約情報を作成したり、あるいは複数項目を統合してまとめや解説情報へと加工したりといった、情報の再構成や表現形態の変更が可能となる。しかし、既存情報の多くは、人間が情報を知覚する際にはその構造を認識することはできても、内部的には構造表現を持っていない。したがって、表現としては平坦となってしまった情報を構造化する技術が必要となる。

マルチメディア情報を構造化する手法として、AV (Audio Visual) 情報と同期しているテキスト情報を用い、自然言語処理によってそのテキスト情報を構造化することにより、副次的にAV情報の構造を得る手法が提案されている[1]。このようなテキスト情報は、例えば字幕放送 (closed caption) によって得ることができる。しかし、日本ではその普及は低く１週間にわずか16時間しかない。しかもそのほとんどがドラマとアニメで、ニュースに関しては皆無である[2]。一方、音声認識によってこのようなテキスト情報を得ることは、不特定話者によるさまざまな文章に対応しなければならないため、現状では困難と考えられる。したがってAV情報から直接構造を認識するAV情報構造認識技術が必要となる。

従来、動画像処理として、運動物体の抽出とその記述、あるいは背景との関係の表現などの動画像認識／理解、あるいはカット点の自動検出[3]やカメラの移動情報の検出[4]などが研究されてきた。これらはいずれも、検索や編集（オーサリング）支援が目的であり、情報をどう構造化するかは利用者に依存する。一方、人間がマルチメディア情報を受けとる際に知覚する情報構造は、当然音声と画像の相互作用によると考えられる。しかし、マルチメディア処理といいながら、音声処理と画像処理を組み合せた情報構造化技術はほとんどない。すなわち、各メディア処理は、それぞれを単独で用いる以外には利用されてこなかったのが現状である。

動画と音声からなるテレビ番組は時間的に連続した単なるデータに過ぎないが、受け手は話題の変わり目を認識したり、話題間の関係を理解している。その認識には、動画や音声が持っている意味的な内容はもちろんであるが、送り手側による表現上の工夫の影響も大きいと考えられる。例えばTVニュースでは、各ニュース項目の始めにはその項目内容を表現する文字スーパ（見出し）が大きく表示される。受け手は、見出しの意味的な内容を理解しなくても、見出しを見るだけで、すなわち見出しが表示されたことを知るだけで、ニュース項目が変わったことを認識できる。このような関係、すなわち情報の知覚レベルにおける表現と人間の情報構造認識との関係を明らかにできれば、AV情報構造化の有効な指針を得られれると考えられる。

本報告では、TVニュースを対象として、人間の知覚という観点からマルチメディア情報の構造について検討し、その構造化手法について述べる。

2．人間の情報構造認識モデル

2.1 題材の選択

本報告では、以下の理由によりTVニュースを題材として選ぶ。

(1) 手に入りやすい

毎日多くの報道番組が放送されている[5]た

め、AV情報の入手が容易に行える。また、TVニュースに対応する意味情報としての文字情報も容易に入手できる。例えばNHKニュースのニュース項目表や、アナウンサが読んだニュース原稿が、Niftyや日経テレコン等を通じて提供されている。文字情報の入手により、本報告で示すAV情報構造の妥当性を評価できる。

(2) 情報伝達を主目的とする

ニュースの主目的は、（少なくとも建前上は）社会的に重要な事実を効率的にかつ正確に伝達することである。このため、映像表現を駆使した叙情的な表現を用いるというよりはむしろ、誤解を生まず一義的に解釈できるような表現形態となるよう工夫している。

(3) 将来的にも適用可能である

企画・構成ものを除き、取材から放送までの時間が短く、編集時間の短縮化が求められている。もちろん、現場等からの生放送も多い。TVニュースは、今後ますます速報性を重視していくと考えられ、この時間的制約が緩和されることはない。すなわち編集段階で構造情報を埋め込んだり、構造情報を同期して送信することは事実上不可能である。

2.2 情報構造認識モデル

情報の送り手の専門家は、情報の受け手がその内容を容易に認識し理解できるように、さまざまな工夫を行っている。例えば新聞の場合、各記事は「見出し・概要・本記・補足（解説）」という共通の構造化がなされ、主要な点は最初の段落に書かれる、いわゆる逆三角形型の構成となる。この構成はそれぞれの記事の意味的な内容には依存しない論理的な構造で、情報伝達を効果的に行うための一種のプロトコルととらえることができる。これらの論理構造の各要素は、例えば見出しは大きな文字で表示するというように、それぞれ固有のレイアウトスタイルがとられる。さらにケイセンなどを用いて構造自身も可視化されているため、１ページに複数の記事が載っていても読者は容易に各記事を区別したり、記事間の親子関係を認識することができる[6]。逆に言えば、レイアウト構造を解析することによって、各情報の論理構造を抽出することができ、その結果意味解析を行わずに情報の構造化が行えるのである。

AV情報に関しても同様と考えられる。すなわち、送り手が映像制作の専門家である場合には、受け手が容易に内容を認識し理解できるように情報を構造化し、それぞれの構造の要素に固有の（時間的な）レイアウト要素を対応付けて送信している。例えば多くのニュース番組では、ニュース項目の始めにニュースキャスタの映像（＋見出し等）を挿入することにより、それぞれの項目の境界を明確にしている。また、いくつかの項目を短く連続して提示するヘッドラインニュース（ニュースフラッシュなどとも呼ばれる）では、ワイプによる場面切換えがしばしば用いられる。ワイプはもともと劇場の暗転や回り舞台に相当する効果を狙いとしている[7]ため、ニュース項目の境界の表現に使用される。さらに、複数のニュースキャスタを用意しておいて項目ごとに話者を変えたり、「さて」「次に」等のクルーワード（キューフレーズとも呼ばれる）や「ポーン」等の効果音を用いたりというような、聴覚的効果を重畳させている。これらの工夫の結果、受け手は、時間的に連続しているAV情報の中から、話題の転換点や、話題の階層構造などを共通して認識することができる。当然のことながら、その認識された構造は送り手の意図している構造と一致しているはずである。

このように、情報が効果的に送受されるためには、その意味的な内容に依存しない、ある意味で普遍的な論理構造が用いられている。これらの論理構造を総称して情報伝達構造と呼ぶことにする。情報伝達構造はレイア

ウト構造と対応して可視化されている。したがってレイアウト構造から逆に情報伝達構造を計算し、それによって情報を構造化すれば、有用なマルチメディア情報を作成できると考えられる。

人間がAV情報から話題の転換点を認識する場合、たとえばカット点や話者の変り目等、視覚的・聴覚的な不連続に起因すると考えられる。換言すれば、送り手の専門家はこの不連続性をうまく利用して情報伝達構造を作り出しているのである。したがって、人間の情報構造認識は、これら不連続点の重み付き総和としてモデル化できる。すなわち、人間に不連続を認識させるメディアの数をnとし、各メディアiにおいて不連続認識に影響を与える要因の数をm_iとしたとき、各時点sにおけるAV情報の切れ目の強さz_sによって以下のように定義できる。

$$z_s = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \delta_{ij} \cdot x_{ij} \qquad (s=1,2,\ldots) \qquad \cdots 式1$$

ただし、各メディアのそれぞれの要因の重み（影響の強さ）をx_{ij}とする。また、δ_{ij}は、sにおいて、各要因に該当する場合には1、非該当の場合は0をとる変数である。

3. 人間の情報構造認識

3.1 実験

TVニュースは、様々な部分から成り立っている構造化されたAV情報ととらえられる。そこで、これらの部分を人間がどのような単位で認識しているか、またそれはどのような要因によるのかを明らかにし、人間の情報構造認識に関するモデルを検証する。

まず、予備実験としてTVニュースを何本か見てもらった。各被験者は、それを意味的な単位に分割し、分割した場所のタイムコードと、分割されたそれぞれに対応する内容を簡単に記述した。その結果以下の2点が明らかとなった。

(1) 分割点の精度を数秒と見たとき、すべての人が分割する時点が存在する。

(2) (1)の時点によってもとのTVニュースを分割したところ、そのTVニュースの項目表の最も大きな単位と一致する。

(1)より、受け手は、時間的に連続しているAV情報に対し、話題の転換点や、話題の階層構造などの認識において共通点があることが示された。また(2)より、共通して認識された構造は送り手の意図している構造、すなわちニュース項目表と一致していることが確認され、情報伝達構造の存在を検証できた。

次に、各個人によって分割の仕方に差が出る部分、すなわち上記予備実験において個人間でばらつきの出た分割点の特徴を調べるため、以下の実験を行った。

各被験者は、家庭用のビデオデッキをジョグシャトル付きのリモコンで操作しながら、一連の関連情報からなるTVニュースの1項目分（2～5分）を見て、その意味的な内容をもとに細かい単位に分割し、スクリプトシートに記入する。スクリプトシートには、分割した場所のタイムコード（画面に表示されている）と、分割されたそれぞれの内容、およびそれらの間の親子構造について簡単に記述してもらった。タイムコードはそれぞれのフレーム画像に書き込まれているため、各試行におけるタイムコードは同一のものが表示される。被験者は、20代から40代の女性20人で、それぞれが5項目について試行した。

3.2 実験結果および解析

実験結果の解析では以下の仮定をおいた。

まず、実験後の被験者の内観に基づいて、動画像の最小単位をショット、また言語の意味的な1まとまりの最小単位を文と考え、以下を仮定する。

仮定１：被験者が記録した区切りは、カット点（動画像）、または文の始まり（音声）を基本とする。

ここでカット点とは、撮影カメラのスイッチのオン／オフ、カメラの切り替わり、編集等によって生じた、動画像内容の時間的あるいは空間的な不連続点のことである。またカット点からカット点までの一連の動画像区間をショットと呼ぶ。被験者は細かなVTR操作に慣れてはいないため、被験者が記録したタイムコードは、実際の動画像／音声の区切りとずれている場合が多い。このため記入されたタイムコードと、カット点または文の始まりとの「ずれ」が１秒前後の場合にはそれを修正する。

一方、TVニュースでは、映像が始まって１秒程度経ってから音声が始まる場合が多い。これを同期して始まったものと見なしても今回の実験目的には影響を与えないと考えられる。したがって、以下の仮定をおく。

仮定２：ショットの始まり（カット点）と文の始まりとが（意味的に）明らかに対応している場合、１秒前後のずれは同期して始まったものと見なす。

仮定２は、被験者が記述したタイムコードの精度、およびニュースの作成／送出過程を考慮すれば妥当である。

さらに、前章で検討した人間の情報構造認識に関するモデルにおいて、情報構造認識に影響を与えるメディア（以下では説明変数と呼ぶ）、および各メディアにおける要因（以下ではカテゴリと呼ぶ）として以下を考えることとする。

仮定３：式１においてメディアの数を視覚系と聴覚系の２つとする。視覚系では、各ショットをキャスタショット・動画ショット・静止画ショット（フリップを含む）の３種類に分け、それぞれの連続関係により以下の６つのカテゴリに分ける。

C_{11}: ショットの変り目なし（連続映像）
C_{12}: キャスタショット ⇔ 静止画ショット
C_{13}: キャスタショット ⇔ 動画ショット
C_{14}: 動画ショット ⇔ 動画ショット
C_{15}: 動画ショット ⇔ 静止画ショット
C_{16}: 静止画ショット ⇔ 静止画ショット

一方聴覚系では、言語的変数（文の変りめとクルーワード）と音響的変数（話者の交替）とに分けて考え、それぞれの有無によって以下の５つのカテゴリに分ける。

C_{21}: 連続
C_{22}: クルーワードなし，話者交替なし
C_{23}: クルーワードなし，話者交替あり
C_{24}: クルーワードあり，話者交替なし
C_{25}: クルーワードあり，話者交替あり

上記以外の因子としては、視覚系では文字スーパのインとアウト，文字スーパの大きさ，映像効果（ワイプ等）が、また音声系では間（ポーズ）の長さなどが考えられる。今回の実験では、特に画面全体の画像としての情報が与える影響の強さについての知見を得ることとを目的とし、このため文字スーパの影響に関しての詳細な検討は省いた。またワイプに関しては種類が多いことから今回の解析では省くこととした。一方、間の長さに関しては、今回の実験精度および仮定１，２を考慮して検討からは除いた。

これらの仮定に基づいて実験結果を整理したものを表3.1に示す。表3.1において、サンプルとは仮定３の説明変数に基いて抽出された時点である。AV情報とは実験データとして用いたTVニュースのスクリプトで、表3.1では１分34秒, １分46秒,…の時点でカット点があり、１分34秒まではキャスタショット、それ以外は動画ショットであることを示している。また、１分20秒, １分34秒,…の時点で音声（文）が始まり、１分34秒の時点では話者交替も

表3.1 実験結果

サンプル	AV情報 タイムコード	視覚	聴覚	説明変数 視覚	説明変数 聴覚	区切りの判断 被験者 1	2	3	…	20	判断率 (%)
1	01:20--	キャスタ	文	C11	C22	1	0	1		0	25.0
2	01:34--	動画	文	C13	C23	0	1	1		0	75.0
3	01:46--	動画	話者	C14	C21	0	0	0		0	0.0
4	01:58--	動画	文	C14	C22	1	0	0		1	10.0
5	02:04--			C14	C21	0	1	0		0	5.0
6	02:09--	動画	文	C11	C22	0	1	0		0	10.0
7	02:14--	動画		C14	C21	1	0	0		0	5.0

あったことを示す。各被験者がそれぞれのサンプル点を区切りと判断したかどうかをそれぞれ1, 0で示し、区切りと判断した人の割合が判断率である。

さて、仮定3により、式1の、説明変数の数n＝2、視覚系・聴覚系のカテゴリ数m_1, m_2はそれぞれ6，5である。そこで、各時点sにおいて予測値z_sと実際の判断率との誤差が最も小さくなるようにカテゴリ数量x_{ij}を求める。すなわち判断率を外的基準として、数量化理論第Ⅰ類によって解析する。解析結果を表3.2に示す。表3.2から、以下のことがわかる。

(1) 外的基準（判断率）の観測値と予測値の重相関係数Rにより、分析結果の精度がわかる。この場合R＝0.89であり、結果の精度は良好である。

(2) 判断率の変動のうち、ここで示した視覚系と聴覚系の2つにより、およそ80%（決定係数R^2=0.80）が説明されている。

(3) 各説明変数のカテゴリ数量の範囲から、外的基準（判断率）への影響の度合がわかる。この場合、聴覚系＞視覚系となることから、区切りの判断は聴覚系の説明変数によるところが大きいことがわかる。

(4) 各説明変数の各カテゴリ数量から、外的基準（判断率）への影響の仕方がわかる。視覚系では、キャスタショットと他のショットとの接合時（C_{12}, C_{13}）に区切りと判断しやすく、他のカット点や連続している場合には区切りとの判断をしにくいことがわかる。一方、聴覚系では、文が変っただけ（C_{22}）では特には区切りと判断せず、クルーワード（C_{24}, C_{25}）や話者交替（C_{23}, C_{25}）と併用されることによって区切りと判断されやすいことがわかる。

(5) 各説明変数と外的基準（判断率）との純粋な相関は偏相関係数でわかる。この場合、偏相関の高い説明変数の順序は聴覚系→視覚系である。

表3.2 実験結果の解析

説明変数	カテゴリ	頻度	カテゴリ数量	基準化カテゴリ数量	範囲	偏相関係数
視覚系	C_{11}	35	x_{11}: -13.916	-12.90	33.67	0.61
	C_{12}	4	x_{12}: 14.409	15.42		
	C_{13}	13	x_{13}: 19.754	20.77		
	C_{14}	50	x_{14}: 0.029	1.04		
	C_{15}	5	x_{15}: 9.145	10.16		
	C_{16}	3	x_{16}: 4.720	5.73		
聴覚系	C_{21}	38	x_{21}: 0	-22.44	84.19	0.85
	C_{22}	51	x_{22}: 21.868	-0.57		
	C_{23}	19	x_{23}: 55.450	33.01		
	C_{24}	11	x_{24}: 69.977	47.59		
	C_{25}	1	x_{25}: 84.188	61.75		

重相関係数R ＝ 0.89　　決定係数R^2 ＝ 0.80

3.3 考察

3.2で示した予測値zは、AV情報の各時点における「区切りの強さ」を表していると考えられる。したがって、zを0から1の範囲に正規化した値によって、AV情報の各時点における「区切りの強さ」を定義する。

「区切りの強さ」により、各ニュース項目を構造化できる。一例として、実験で用いた「公定歩合引き上げ」のニュースに対して計算された区切りの強さを図3.1に示す。図3.1から明らかなように、多くの人が区切りと感じた点、すなわち「区切りの強さ」が大きい点

と、項目表から得られる項目内構造とは一致し、予備実験で示されたよりももっと細かい単位での情報伝達構造の存在、および情報構造認識モデルが正しいことが検証できた。したがって、TVニュースの構造化では、特にキャスタショットの検出と、話者交替点の認識およびクルーワード検出が重要であることがわかる。なお、ここではキャスタショットという語を用いたが、AV情報全体にわたってその進行を制御している人物または事物のショット、例えばのど自慢やクイズ番組の司会者、教養／討論番組の進行役や議長役の人物、あるいは野球中継やダイジェストのスコアボードととらえることにより、さまざまなAV情報に対応できると考えられる。

以上の検討結果に基づき、進行制御人物ショット検出技術、および話者交替検出技術を開発した。開発されたアルゴリズムでは、登場するニュースキャスタや話者に関する事前の学習を必要とせず、話者数などの情報も用いないため、幅広いソースに対応できる。実現されたプログラムは、それぞれ90%以上の認識精度を得ている。またNHKニュースを対象とした情報構造認識結果は、ニュース項目表と85%程度の精度で対応付けられた。今後、これ以外のソースに対する評価を行う。

4．情報要約への応用

AV情報を構造化する利点の1つは、受け手側の制約（可用な時間やメディア、あるいは各ユーザの好み等）に応じた情報提示が可能となることである。その一例として要約情報の作成／提示について述べる。

TVニュースは、通常、はじめにニュースキャスタが登場してそのニュース項目の概要を喋り、その後、現場の映像と共に詳細な内容を音声によって伝える、という構成となっている。この構成は情報伝達の基本構造であり、受け手もこのような構造を認識していることは3章で示した実験によって確かめられている。このうち、はじめの概要部分のみを取り出せば、いわゆるヘッドラインニュースができる。通常ニュースの1項目は1～2分前後だが、そのうちのリード部分は10～15秒程度である。したがって、10項目からなるニュースならば2～3分にまとめることがで

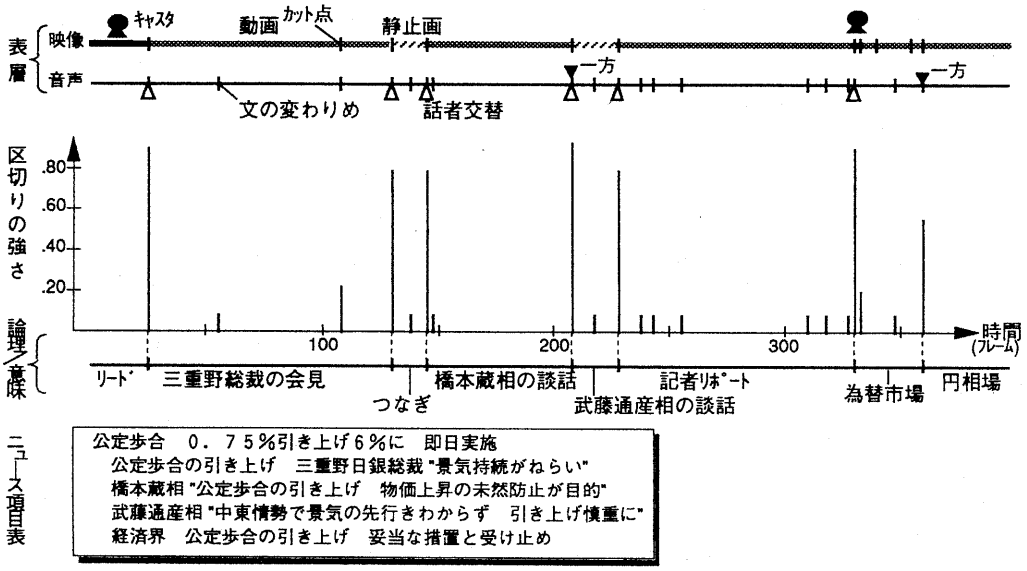

図3.1 区切りの強さによって抽出されたTVニュースの構造とニュース項目表との対応

きる。また、生放送のニュースは、常に時間の調整をしながら進行するため、編集した映像が全部放送されるとは限らない。このため、ニュースを最も象徴する映像や重要な意味のある映像は編集の前段に配置される[8]。そこで、概要音声と、その長さに合わせて切り出された詳細映像とを同期させて提示すれば、より効果的な要約情報となる（図4.1）。

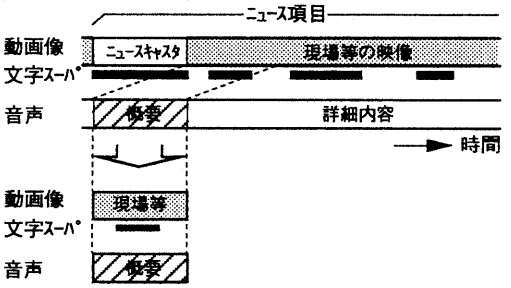

図4.1 TVニュースの要約作成

あるいは、重要な情報は正確に伝わるように文字スーパが用いられると仮定すれば、文字スーパの提示されている部分を少しずつ切り出して概要音声の長さに合うようにして提示する方法も考えられる。後二者は音声をもとに動画像を再編集し提示することとなるが、予備実験では特に違和感なく受け入れられた。今後詳しい実験によって検討を加える予定である。

このように、送信者の意図、および受信者の情報取得を念頭において情報を構造化しておくことにより、効果的な情報編集／提示が容易に行える。

5．おわりに

大容量通信を可能とするインフラの整備、圧縮技術の発達とLSI化、記憶装置の低価格化と大容量化、動画や音声を扱えるパソコンの登場、などマルチメディア情報の伝送／蓄積／提示を可能とする環境は整いつつある。これらハード先行型で進んできた「マルチメディア技術」ではあるが、いったいその上でどんな情報を流すの、それを誰が入れるのといった問いへの答はまだない。しかし、βやキャプテンを引合に出すまでもなく、結局大切なのはソフトである。すなわち、今後「マルチメディア」が発展するかどうかは、そこに流すべき情報をどのような指針で（自動的に）構造化し入力できるのかが、大きな鍵を握っていると考える。

本報告で示した情報伝達構造は、情報の送受信者間に介在する暗黙の論理構造であり、マルチメディア情報を構造化する際の有効な指針となり得ると考える。さらにこの構造は、提案した情報構造認識モデルに基づいてAV情報から直接計算することができる。

今後、処理時間や運用コストを念頭においてフィージビリティを検討するとともに、サービスとして具体化したときの問題点について検討していきたい。

参考文献

[1] 竹下：文タイプ情報を用いた話題構造の認識，AI学会SIG-SLUD-9203-2, 1992, 9-18.

[2] 「日本のTV字幕放送　欧米に比べ障害者にお粗末　文章変換に手間と費用」，読売新聞朝刊, 1993.6.25.

[3] 大辻他：動画カット検出，信学技報IE91-116, 1991.

[4] 阿久津他：動画像インテクシングを目的としたカメラ操作の規定方法，信学論D-II, vol.J75-D-II, no.2, 1992, 226-235.

[5] 情報通信年鑑'91，情報通信総合研究所, 1991.

[6] 大久保・小林：一覧性に着目した情報提示方式の検討，信学技法HC91-17, 1997, 47-54.

[7] 小林・岩本：ドラマ番組の編集技術，TV学会誌, vol.44, no.6, 1990, 659-665.

[8] 田村：ニュース番組の編集技術，TV学会誌, vol.44, no.6, 1990, 674-677.