ニュース映像からの重要セグメント抽出
— 画像特徴と言語特徴の相互関係を用いたニュース映像要約

† 中村 裕一
† 筑波大学 電子情報工学系
〒305 つくば市 天王台 1–1–1
(yuichi@is.tsukuba.ac.jp)

‡ 金出 武雄
‡ カーネギーメロン大学
ロボティクス研究所

あらまし: 本論文では, ニュース映像を構成する重要部分を画像特徴と言語特徴の両方を用いて抽出する方法を提案する. 具体的には, 画像として顔, 人物像, 屋外風景等を抽出し, 自然言語 (transcript) からは, 発話, 会合その他の特徴的な部分を抽出する. これらを照合することにより, ニュース映像で重要な意味を持つ部分を抽出する. 抽出されたセグメントを提示することにより, トピックのわかりやすい要約ができあがる. 本研究では米国 CNN ヘッドラインニュースについて実験を行い, 良好な結果を得た.

キーワード: 映像処理, 映像要約, ニュース映像処理, 画像処理と自然言語処理の統合

Content Extraction from News Video
— Video Summerization by Image and Language Analysis

† Yuichi NAKAMURA, ‡ Takeo Kanade

† Institute of Information Sciences and Electronics, University of Tsukuba
‡ The Robotics Institute, Carnegie Mellon University

Abstract: Spotting by Association method for video analysis is a novel method to detect video segments with typical semantics. Video data contains various kinds of information through continuous images, natural language, and sound. For videos to be stored and retrieved in a Digital Library, it is essential to segment the video data into meaningful pieces. To detect meaningful segments, we need to identify the segment in each modality (video, language, and sound) that corresponds to the same story. For this purpose, we propose a new method for making correspondences between image-clues detected by image analysis and language-clues detected by natural language analysis. As a result, relevant video segments with sufficient information from every modality are obtained. We applied our method to closed-captioned CNN Headline News. Video segments with important events, such as a public speech, meeting, or visit, are detected fairly well.

Keywords: news video analysis, content extraction, spotting by association, video summarization

1 はじめに

Informedia プロジェクト [WKSS96] では, ニュース映像やドキュメンタリ映像を大量に蓄積して, その検索, 要約などの研究を行っている. 例えば, ニュース映像の検索システムでは, ユーザの音声/タイプ入力 (例えば, “tell me about mad cow disease”) によって, それに関連するニューストピックを検索することができる.

大量の映像データの中から, 簡単な質問によって映像を検索するシステムでは, 効率的な検索と共に, 検索されたデータのわかりやすい提示方法が重要になってくる. 例えば, 1995-96 年にかけてのニュース映像からクリントン大統領がアイルランドで演説したシーンを探すのに, “Mr. Clinton” と “Ireland” を検索したとすると, かなりの数の関係の薄いデータが検索されるだろう. そこで, 個々のトピックの重要部分を抽出し, 意味的なタグを付けることと, 検索された結果をわかりやすく表示する処理が必要となる.

そのための最も簡単な要約方法の一つは, トピックの内容を良く表す比較的少数の画像と言語 (文又は音声) の組を選び出し, これらの組を使ってトピックの索引付け, トピックの要約/提示を行うことがである. ところが, これまでの研究では, このような組として, トピックの最初の文と最初の画像を用いている場合や, 逆にトピック内の全ての画像と発話内容 (transcript) を用いている場合が多い. これらは必ずしも, トピックの良い説明とはならない. この問題の解決策として, Zhang らはいくつかの画像特徴を使って代表フレーム (key-frame) を選び出す方法を提案している [ZLSW95]. また, Smith らは, TFIDF を用いて提示する単語 (場合に応じて, 句や節) を選択し, カメラワーク等の画像特徴を使って, 各カットの内容を最もよく表現する画像を選択している [SK97]. これらは, トピックの内容を深く処理することなしに, ユーザに提示するデータ量を減らすという点で興味深く, 広く通用する技術である.

しかし, 依然として, いくつかの問題点が残っている. 一つには得られたデータを取捨選択するのが難しいことあげられる. これは, 各々のセグメントが何を説明しているかを考えていないからである. もう一つは, 画像と言語の対応関係がとれているかどうかの確認ができないことがある. 画像と言語の各々が異なることを説明している場合, 誤解を与えるようなトピックの説明になる可能性が大きい.

そこで, この研究では, 画像と言語の表層的な意味なカテゴリを考え, それらが一致している部分を対応づけることによって, 重要なセグメントを選び出す方法 (Spotting by Association) について提案する. つまり, 画像と言語が同期して, 同一の対象について説明している部分を選ぶことによって, 製作者の最も伝えたい部分であり, しかも, 単独で見せられてもわかりやすい部分を抽出することを目的としている.

以下本稿では, ここで提案する手法の考え方, 処理, 実際のニュース映像に適用した結果を報告する.

2 対応づけに基づくスポッティング

2.1 映像における画像と言語の役割

我々は, 画像のみ, 言語のみからでも, ある程度映像の内容を理解できる. これは, 画像, 言語がある程度重複して情報を伝えるからである. 例えば, 図 1(a) のように, 顔のクローズアップがあり, 口が動いていれば, 音声が無くても発言に焦点があてられていることがわかる. 同様に, 車が炎上している場面 (b) を見れば, 交通事故があったこと, 被害の程度がどのようなものであったかを (必ずしも正しくはないが) 推測できる[1].

ただし, 画像と言語は相補的なモダリティであるため, どちらか単独では誤解を招くような場合もある. 例えば, 講演, 意見などに焦点があてられている部分を取り出すことを考えてみよう. 顔のクローズアップをとり出すことは, 現在の CV 技術である程度可能であり, 多くの場合は図 1(a) のように, 正しい結果が得られる. しかし, 図 1(c) のように, 発話のシーンではなく, 事件の被害者の説明として, 顔のクローズアップが出てくる場合もあり得る. 同様に, 誰かが意見を主張している部分を言語的説明 (transcript) から抽出することを考えてみよう. 人間が十分に考えたうえで抽出を行うと, おそらくほとんど正しいものが得られるが, 現在の計算機の自然言語処理技術では, 完全なものは望めない. 例えば, “They say ...” というのが本当の発言なのか, ただの噂話なのかを判断することは計算機にとって難しい. そこで, 両者から推定される状況が一致する部分を取り出すことが有効となる.

2.2 Spotting by Association

本研究では, 言語と画像を関係づけることにより重要セグメントを抽出する方法 (Spotting by Association) を提案する. この方法では, 画像と言語の意味的なカテゴリを考え, それらが一致している部分を対応づける. このように, 言語と画像が同一対象について矛盾しない情報を与えている部分を抽出することには 2 つの利点がある. 一つは, 画像と言語の両方を使うことによって, 処理がより正確になるという点である. もう一点は言語と画像両方が矛盾なく揃っているデータ, つまり, 他の部分と独立して提示されても分かりやすいデータを得ることができるという点である.

本研究では, 対象がニュース番組であるため, ニュース番組におけるいくつかの典型的な状況を意味的なカテゴリとして選び出した. これらは, 図 2 に示すような, 発言, 会合, 訪問, デモンストレーション, 風景による現場説明である.

[1] 実際にはミサイル爆撃を受けて炎上した車の映像である.

(a) (b) (c) (d)

図 1: Example of images in news videos

表 1: Clues from language and image

language clues	
SPEECH OPINION	speech, lecture, opinion, etc.
MEETING CONFERENCE	conference, congress, etc.
CROWD PEOPLE	gathering people, demonstration, etc.
VISIT/TRAVEL	VIP's visit, etc.
LOCATION	explanation for location, city, country, or natural phenomena
image clues	
FACE	human face close-up (not too small)
PEOPLE	more than one person, faces or human figures
OUTDOOR-SCENE	outdoor scene regardless of natural or artificial.

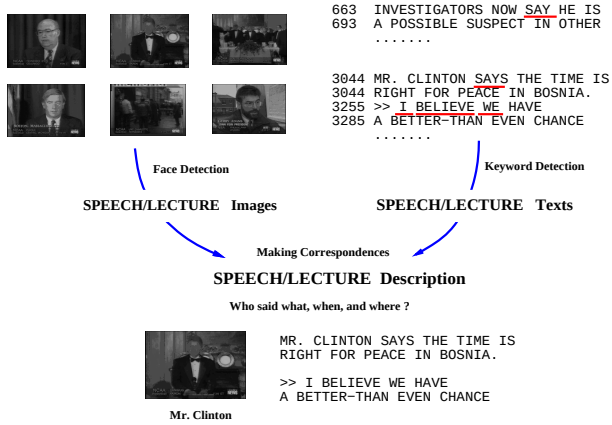


図 3: Basic idea of Spotting by Association

これらの状況を抽出するための大まかな考え方を 図 3 に示す．基本的には，特定の状況を表し得る画像 (表 1) を各カットの代表フレーム (poster frame) から選びだし，それを key-image とする．同様に，ある状況を表し得る特徴的な文 (表 1) をクローズトキャプション (transcript) から抽出し，key-sentence とする．次に，これら 2 つの時系列のデータを動的計画法 (DP) によって関係付けることによって，意味的に整合性のある言語と画像の対が得られる．この方法により対応のとれた部分が映像において重要な断片だと考える．図 3 はスピーチのみを抽出する場合の例であり，key-image として顔画像を抽出し，発言を示唆する部分をクローズトキャプションから抽出することになる．

3 言語からの特徴抽出

クローズトキャプション (closed-caption) はニュース番組の映像信号から自動的に計算機に取り込まれる．したがって，映像を MPEG にエンコーディングする際のフレーム番号と大まかな対応がとれる．

これによって得られた発話，ナレーションデータから，上記の *key-sentence* を抽出する．その一番簡単な方法はキーワードスポッティングであるが，かなり多くの無駄な単語が抽出されるため，補助的に構文解析を行ない，不要なキーワードを除去する．

3.1 キーワードスポッティング

特定の状況を推測させる部分を抽出するためには，キーワードスポッティングが最も簡単な手法である．本研究で設定した状況においても，特定のキーワードが頻繁に現れる．例えば，図 3 に示されるように，発話を示唆する文章には以下のような単語が現れることが多い．

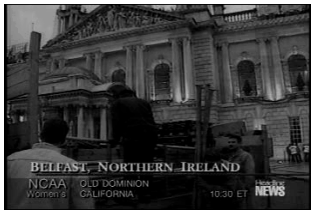
発話について言及: say, talk, tell, claim, acknowledge, agree, express, etc.

話者の発話内: I, my, me, we, our, us, think, believe, etc.

一番目は，レポータやアンカーパーソンが他人の発言に対して言及するために良く使われる単語である．CNN Headline News ではこれらの単語が，発話の説明部分の大勢を占める．二番目は，発言者自信の発話内に現れ易い単語の例である．これを実際に割合として示したのが 表 2 である．表中の数字は，スピーチを表す状況において，各々の単語が現れた割合である．ただし，未来形，否定形の文で現れる場合は，実際に起きたことがらではない場合がほとんどであるため，ここでは除いている．この表を見ると，“say”，“I” 等は非常に高い度合でスピーチを表すシーンと同期して現れている．同様に，会議，訪問の状況についての例を表 3 に示す．

3.2 構文解析によるキーワード選択

表 2 を見ればわかるように，“talk” 等はスピーチシーンを抽出するための十分な手がかりとはならない．その一つの大きな原因は，“peace talk” のように，名詞形で用いられることが多く，話すという動作よりも，話のトピックに焦点が当たっている場合があるからである．また，否定形，未来形の文に表現されている事柄が，実際に画像とし



SPEECH/OPINION & FACE

Who spoke what?
Where?

MEETING/CONFERENCE & PEOPLE

Who met whom?
What subject?

OUTDOOR SCENE & LOCATION

Where?
What event?
Who visit where?

図 2: Typical situations

表 2: Keyword usage for speech

Indirect Narration

word	speech	not speech	rate
say	118	11	92%
tell	28	3	90%
claim	12	6	67%
talk	15	37	29%

Direct Narration or Live Video

word	speech	not speech	rate
I (my, me)	132	16	89%
we (our, us)	109	37	75%
think	74	15	84%
believe	12	10	55%

表 3: Keyword usage for meeting and visiting

word	human meet	others	rate
meet	31	9	78%
see	15	59	20%
word	**human visit**	**others**	**rate**
visit	21	1	95%
come	30	62	32%

て与えられることは非常に少ない [2]. そのため, キーワード単独でスポッティングを行うと, 多くの誤った候補が抽出される. ある程度の抽出誤りは, 画像との対応関係を調べることによって落すことができるが, 大量に誤りが出てきた場合には, それは難しい.

そのため, キャプション文を構文解析し, 主語, 動詞, 目的語といった各単語の役割を抽出する. これによって, 以下の 4 つのことが可能になる.

1. 使われている品詞によって, キーワードとしての重要性を判定する. 例えば, “talk” が名詞で主語である場合と, 動詞である場合の区別をつける.
2. 単語の主語 (動作主) を調べることによって, キーワードとして適切かどうかを判定する. 例えば, “talk” が動詞である場合には, その主語が人間であるかどうかを確かめることによって, 実際の発話かどうかを確かめることができる. そのためには, 主語の上位概念に人間を表す単語があるかどうかを調べれば良い. この処理は, WordNet[Mil90] を用いて行う.
3. 文が否定形, 未来形かどうかを調べ, その場合には候補から除く.
4. “in”, “to”, “from” 等の前置詞と場所を表す可能性のある単語の組み合わせを調べることにより, 場所について言及されている部分をより確実に抽出することができる.

[2] ドラマや SF のように, 仮想的な状況を映像にする場合にのみ現れる.

表 4: Conditions for *key-sentence* detection

type	condition
SPEECH OPINION	active voice and affirmative, not future tense, subject as a human or a social group, not “it”
MEETING CONFERENCE	affirmative, not future tense
CROWD	affirmative, not future tense
VISIT TRAVEL	affirmative, not future tense, subject as human, at least one location name in a sentence
LOCATION	preposition (in, at, on, to, etc.) + location name

3.3 処理過程

処理は以下のようになる. まず, Link Parser によって構文解析を行う [ST93]. ニュース映像の文章は比較的複雑で, 現在の自然言語処理技術では構文解析結果の誤りが多い. しかし, 主語と動詞だけに注目した場合には, 8 割程度の精度が得られる. これを用いて, キーワードの品詞, 動詞の場合の主語を知ることができる.

次に, 主語及び文中の単語の意味を調べる. 調べるべき単語と, その条件は表 4 に示す通りである. チェックに残った単語を含む文を key-sentence とする. その結果, 表 5 のような結果が得られる. 表の各項目 (X/Y/Z) の X はキーワードが存在する文の数, Y は上記の処理によって除かれた数である. Z は除去されたものに, 正しい key-sentence が含まれていた数である.

4 画像からの特徴抽出

ニュース映像で非常に大きな意味を持っているのが人の行動であるため, 必然的に, 人物の映像が重要な意味を持つ. また, 人の移動や訪問, 自然現象に関して, 場所, 特に

表 5: Key-sentence detection result

	speech	meeting	crowd	visit	location
Video1	40/3/1	20/1/0	33/4/0	41/33/0	89/59/5
Video2	28/3/0	22/6/0	24/3/0	39/34/1	65/39/2
Video3	34/5/1	15/2/1	22/2/0	39/33/0	70/50/4

表 6: Usage of face close-up

video	speech	others	total
Video1	59	10	69
Video2	80	12	92

Other usages are personal introduction(4), action(2), audience/attendee(3), movie(2), anonymous(2), exercising(2), sports(1), and singing(4).

屋外風景が重要な意味を持つ.

4.1 Key-image

本研究では, 顔 (クローズアップ), 複数人物, 屋外風景の 3 つを key-image とする. ここで, 3 つの種類の画像がどのような目的で用いられているかを調べた結果を 表 6–表 8 に示す.

顔のクローズアップは, 発話, 人物の紹介等の目的で用いられることが多い. また, 複数の人物が同時に映っている場合の画像を図 4 に示す. このような場合には, 表 7 に示されているように, 何らかの目的のために集まっている人々を表す場合が多いが, その中でも, 国会や裁判などの会議の場面が多い. 典型的な場合には, 会議の名前だけが言及される. その他の場合には, “people” 等, 人々を表す単語で説明されていることが多い. 屋外風景の例を 図 5 に示す. 屋外風景は, イベント, 事件, 自然災害に関するトピックでは重要な意味を持つ. それらは単純な場所の説明から, 災害の程度を表現したりする. これらがニュース映像で果たす目的は, 土地, 町の漠然とした状況の説明, 天候の説明, 建物によって企業や組織を表現する, 等である. ただし, これらの目的を厳密に切り分けるのは難しいので, 表 8 に全体の数だけをあげる.

これらの性質を考えると, 単独の画像だけで, 画像が映像中のどの状況を表すかを推定するのは難しいが, 典型的ないくつかの候補に分類することが可能である. そこで, 言語的説明との対応関係を用いることによって, その画像が表す状況をより確実に推定できる.

4.2 Key-image の抽出

まず, 前処理として, ヒストグラムを用いたシーンチェンジ抽出方法を用いて映像をカット分割する. 各カットの最初の (厳密には 10 フレーム後) をそのカットの代表フレームとする [SH95, HS95]. 次に, 各代表フレームについて, 以下のような特徴抽出を行う.

顔のクローズアップの抽出: 各代表フレームに Rawly らの顔検出プログラムを適用する [RBK96]. 基本的に, 大きく, 中心にあるほど顔のクローズアップであると考えられるため, 抽出は比較的簡単である. 我々の実験では, 顔のクローズアップの 8 割以上が抽出されている.

(a) (b)

図 4: Example of people images

表 7: Usage of people images

video	meeting	crowd	total
Video1	16	16	32
Video2	9	43	52

人物, 屋外風景画像の抽出: 会議や人物集団を表す画像の多くには, 小さな顔や人物像が含まれる. そこで, 顔のクローズアップと同様の方法 [RBK96] で抽出を試みると, 実際に小さな顔の含まれる画像の半数以下しか検出されない. これは, 顔が小さくなると, 顔としての特徴が失われるために, 上記の方法では検出の精度が悪くなるためである. 小さな顔でも抽出され易い手法については現在研究中であり, 以下の実験では, 人手で指定している[3]. 屋外風景についても同様に人手での抽出を行っている.

5 DP による対応づけ

5.1 基本的な考え方

画像からの特徴抽出, 言語からの特徴抽出で各々時系列の key-image と key-sentence が得られる. ここで, 各々の時系列データの対応をつけるため, 動的計画法 (DP) を用いる. そのための前提条件としては, 言語で述べることがらと画像で述べることがらの順序が一致することを仮定している. また, 一つのデータが複数の意味的なカテゴリを持つ場合, その数だけ異なるデータがあるものとして扱う. これによって, 一対多の対応を考える必要がなくなり, 簡単な DP の計算で対応を求めることができる.

基本的には, 以下の式のペナルティ (P) を最小化する.

$$P = \sum_{j \in Sn} Skip_{\rm s}(j) + \sum_{k \in In} Skip_{\rm i}(k) + \sum_{j \in S, k \in I} Match(j,k)$$

ここで, S, I は各々対応の見つかった文, 画像で, Sn, In は各々対応の見つからなかった文, 画像である. $Skip_{\rm i}$ は, ある画像 (i) が文と対応しない (スキップされる) ことのペナルティ, $Skip_{\rm s}$ はある文 (j) が画像と対応しないことのペナルティ, $Match(i,j)$ は, ある画像 (i) と文 (j) の対応に対する評価値 (ペナルティ) である.

[3] この作業は比較的簡単なものであり, 30 分のニュース番組に対して 5 分程度の手間となる.

(a) (b)

図 5: Example of outdoor scenes

表 8: Usage of outdoor scenes

video	outdoor scenes
Video1	34
Video2	39

対応に関しては, 時間的に大きく離れ過ぎていない限り, 任意の対応関係を許す方法をとった. つまり, ある画像の生起時間と一定時間以内に起こった全ての文を候補として考える. 本研究の実験ではこの対応窓の広さを 20 秒程度としている. また, 任意の画像, 文をスキップ像 (または文) をスキップすることが可能である.

5.2 コストの評価

Skip のペナルティ ($Skip$): 以下に $Skip$, $Match$ の評価方法を示すが, 基本的に複数の候補の中から正しい候補を選ぶ処理であるため, 絶対的な値は意味を持たない. そのため, 個々のデータがどの程度言語画像間の対応関係を持ち得るかを評価し, それを基にペナルティの値を経験的に決めることになる.

まず, この研究では, 個々のデータの評価 (E_{type}) とタイプによる評価 (E_{data}) の掛け合わせであると考える.

$$E = E_{\text{type}} \cdot E_{\text{data}}$$

例えば "顔 (f_i)" が抽出された場合の画像の評価は以下の式で表される.

$$E = E_{\text{type}}(FACE) \cdot E_{\text{face}}(f_i)$$

E_{type} の具体的な値を表 9 に示す. $E_{\text{face}}(f_i)$ は, その画像がどの程度顔の close-up らしいかの評価値である. 顔の場合には, 顔が画像に占める画素の評価値を足し合わせたものになっている. 同様に, key-sentence の評価方法を表 10 に示す. 各 key-sentence は, 品詞, 構文的な役割, 主語の種類等によって評価される.

マッチングペナルティ ($Match$): key-image と key-sentence の対応関係の評価は, 以下の式になる.

$$Match(i,j) = M_{\text{time}}(i,j) \cdot M_{\text{type}}(i,j)$$

ここで, M_{time} は画像と文の時間的な整合性. M_{type} は画像と文のタイプの適合性を表す. d_i と d_s が一致するほど評価が良くなる (ペナルティが小さくなる).

表 9: Example of cost definition

***key-sentence*:** speech 1.0, meeting 0.6, crowd 0.6, travel/visit 0.6, location 0.6
***key-image*:** face 1.0, people 0.6, scene 0.6

表 10: Example of sentence cost definition

1.SPEECH/OPINION
keyword's part-of-speech: verb 1.0, noun 0.6
subject type: a proper noun suggesting a human or a social group 1.0, a common noun suggesting a human or a social group 0.8, other nouns 0.3
2.MEETING
keyword's part-of-speech: verb 1.0, noun 0.6
subject type: a proper noun suggesting a human or a social group 1.0, a common noun suggesting a human or a social group 0.8, other nouns 0.3
verb semantics: verbs suggesting attendance 1.0, the other verbs 0.8

M_{time} は以下のように決める. まず, 画像に対してはカットの持続時間 (d_i) を, 文に対してはその文が発話されている時間 (d_s) を求める. M_{type} の例を表 11 に示す. これは, E_{type} と同様に, 実際の対応関係を調べ, 適当な値を経験的に決めた. また, 現在の段階では用いていないが, 顔と名前のように, 何らかの既知情報があれば, それを組み入れることが, 精度の向上につながる.

6 実験例

実験に用いる映像として, CNN Headline News の映像 6 本 (一本あたり 30 分, 計 3 時間) を選んだ. 上記で述べた処理を順に行うが, 複数人物検出, 室外風景抽出については, まだ十分な精度が得られていないため, 人手による. それ以外の部分は自動的に処理される.

ある一本のニュース映像 (Video1) を処理した結果, key-image 167 個, key-sentence 122 文が得られ, 対応のとれた数が 69, 対応のとれなかった数が 53 であった. 図 6 は DP によって得られた対応関係の例である.

6 本のニュース映像を処理した結果を表 12, 表 13 に示す. 30 分のニュース番組一本当たり, 約 70 のセグメント (key-image と key-sentence の対) が抽出され, そのうちの約 50 本が正しい結果となっている. また, 対応の見つからなかったものの大部分は, 実際に対応する key-image または key-sentence が無かったものである. 特に, CM 部分を除く処理を行っていないために, 対応のない key-image が多く残っている. また, 対応する部分があるのに間違った対応が得られる場合, その原因として以下のようなものがあげられる.

- 顔 (key-image) の抽出失敗.
- クローズトキャプションと画像の表示時間ずれ.
- key-sentence の抽出失敗. 人物の紹介のように key-

表 11: Matching evaluation for type combinations

	speech	meeting	crowd	visit	location
face	1.0	0.25	0.25	0.25	0.0
people	0.75	1.0	1.0	0.5	0.5
outdoor scene	0.0	0.25	0.25	1.0	1.0

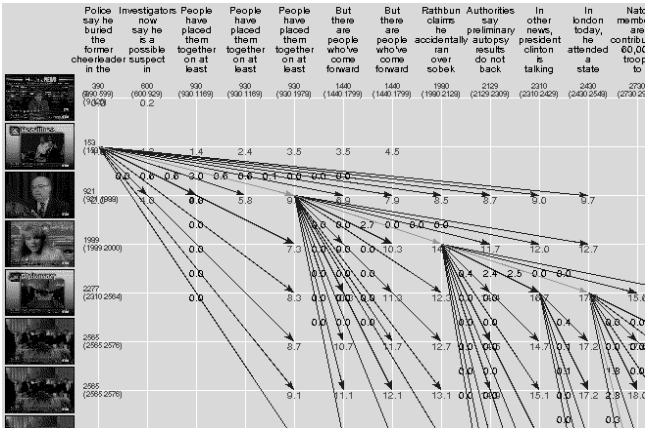


図 6: Correspondence between sentences and images

sentence として取り出しにくい文に対応できていない.

対応があっていても, 話者の顔とそのスピーチのような想定パターンになっていない場合には様々なものがあるが, 良く現れるのは, 以下のような場合である.

- 会議などにおいて, 話者, 聴者等に焦点を当てず, 議会の漠然とした状況を写し続けるような場合.
- 専門家が意見を紹介するような場合には, 話者ではなく話の内容に関連する画像が与えられる場合がある.
- Business 関係のニュースなどで, 会社の説明として会社のある街の風景や建物が出てくるような場合.

結果の利用

得られた key-image と key-sentence の対応関係を使って以下のような映像の再利用方法が考えられる.

要約, 提示: 30 分のニュース映像の実質的な内容は 25 分前後であり, 70 個の重要セグメントが得られたとすると, 1 分間に 3 個弱のセグメントが得られることになる. 画像を静止画, 文をテキストとして表示すれば, 一つの画面に複数個表示できるため, 特に長いものでなければ, トピックの内容を一覧することができる. 実際に状況毎に表示した例を図 7, 図 9 に示す. それぞれの行は場所, 訪問に関する情報, 会議, 群衆に関する情報, スピーチ, 意見に関する情報を表す. 例えば, 図 7 の例では, クリントン大統領の訪問とそれに備えるベルファストの街の様子が一番上の行に, 政治家や民衆の様子が二番目の行に, アイルランドとイギリスの和平に対する各論が三番目の行にあげられている. また, 横軸を時間として, 映像中の時間したがって並べたものを, 図 8 に示す. このように, トピックの意味的構造が一覧でき, 理解しやすい要約表示が可能になる. また, これをブラウジングすることによって, 映像データそのものに簡単にアクセスすることが可能である.

検索: 本研究で得られたデータを基にしたデータのインデッキシング, 検索を考えることができる. ここで提案した手法で完全なデータを得ることは難しいが, 人手の介在を許せば, 比較的少ない手間で修正することができる. また, スピーチの場合の話者, 会議の場合の出席者等の情報を自然言語処理等によって付け加えることができれば, より有用なデータとなるだろう.

表 12: Spotting result 1 (six 30-minute videos)

type	all A	matched B	correct C	miss D	wrong E
speech	292	226	178	40	48
meeting	47	26	19	18	7
crowd	63	35	26	19	9
travel	15	8	7	6	1
location	76	34	27	32	7
face	472	217	173	0	44
people	220	84	63	0	21
scene	168	25	21	0	4

A is the total number of *key-data*, B is the number of *key-data* for which inter-modal correspondences are found, C is the number of *key-data* associated with correct correspondences, D is the number of missing association, that is the number of *clues* for which association is failed in spite of having real correspondences, E is the number of wrong association, *i.e.* mismatching.

表 13: Spotting result 2

	face	people	scene
speech	199/165	24/12	2/1
meeting	9/6	15/12	1/1
crowd	5/1	28/25	1/0
visit	1/0	4/4	3/3
location	3/1	13/10	18/16

Each figure (X/Y) in the following table shows, the number of found correspondences (X) and the number of correct correspondences (Y).

7 まとめ

本論文では, ニュース映像から特別な意味を持った重要な断片を抽出するための方法を提案した. そのための key-image 抽出法, key-sentence 抽出法, 対応の方法について述べた. CNN Headline News にこの手法を適用した結果, 画像と言語の素対応を見つけるという点でかなり良い結果が得られることがわかった. 得られた意味的にも典型的なパターンに当てはまる場合が多く, これを要約, 提示, 検索のためのデータとして使うことが可能である.

図 7: News video TOPIC EXPLAINER (Category)

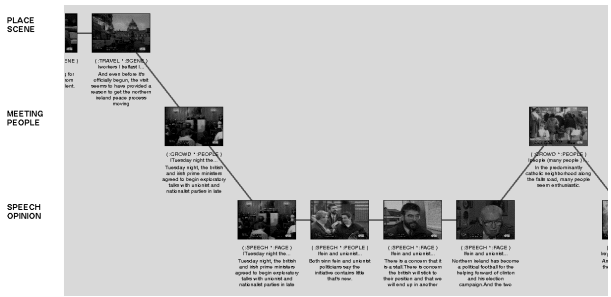


図 8: News video TOPIC EXPLAINER (category + temporal order)

今後は, 自動化を進めるとともに, 多様な入力データに対処することが課題である.

図 9: Details in TOPIC EXPLAINER

参考文献

[HS95] A. Hauptmann and M. Smith. Video Segmentation in the Informedia Project. In *IJCAI-95, Workshop on Intelligent Multimedia Information Retrieval*, 1995.

[Mil90] G. Miller. WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, Vol. 3, No. 4, 1990.

[RBK96] H. Rowley, A. Baluja, and T. Kanade. Neural Network-Based Face Detection. *Image Understanding Workshop*, 1996.

[SH95] M. Smith and A. Hauptmann. Text, Speech, and Vision for Video Segmentation: The Informedia Project. *AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision*, 1995.

[SK97] M. Smith and T. Kanade. Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques. *IEEE CVPR*, 1997.

[ST93] D. Sleator and D. Temperley. Parsing English with a Link Grammar. *Third International Workshop on Parsing Technologies*, 1993.

[WKSS96] H. Wactlar, T. Kanade, M. Smith, and S. Stevens. Intelligent Access to Digital Video: The Informedia Project. *IEEE Computer*, Vol. 29, No. 5, 1996.

[ZLSW95] H. Zhang, C. Low, S. Smoliar, and J. Wu. Video Parsing, Retrieval and Browsing: An Integrated and Content-Based Solution. *Proc. ACM Multimedia*, 1995.