コンピュータビジョンと
イメージメディア 114－18
（1999. 1. 22）

言語情報を伴う画像の画像的特徴量と語義の統計的対応付け

井手一郎　浜田玲子　坂井修一　田中英彦

{ide,reiko,sakai,tanaka}@mtl.t.u-tokyo.ac.jp

東京大学大学院 工学系研究科 電気工学専攻

〒113-8656 東京都 文京区 本郷 7-3-1

増大する映像資源，特に速報性と利用価値の点からニュース映像への自動的索引付けを行う必要が高まっている．これに応えるために，映像に附随する言語情報を利用した自動的索引付けの試みが内外で活発に行われているが，その多くは言語情報主導で映像内容は十分に考慮されていない．このような問題点をふまえ，映像内容を考慮した索引付けを実現するために，あらかじめ教師映像データから映像内容と画像的特徴量の関係を統計的に学習して対応付けておくことにより，索引付けを行いたい未知の映像の画像的特徴量から映像内容を推測する手がかりとしたい．本稿では，このような機構を実現するための手法の紹介と，初期段階の実験結果について報告する．

Statistical Alignment of Textual Semantics with Graphical Features from Images with Captions

Ichiro Ide, Reiko Hamada, Shuichi Sakai
and Hidehiko Tanaka

{ide,reiko,sakai,tanaka}@mtl.t.u-tokyo.ac.jp

Graduate School of Electrical Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656

Automatic indexing to video data, especially to news video, is in strong demand considering its contents' importance and value. Reflecting the demand, various attempts have been made to index news video automatically utilyzing accompanying textual data. Most methods are textual data oriented, which do not consider the video contents thoroughly. Referring this issue, we will propose an indexing method, which considers the video contents, by first aligning textual semantics with graphical features statistically from supervisory video data, and later assuming the image contents of incoming video data from their graphical features. We will introduce the mechanism of this method and report the result of an early stage experiment.

1 序論

1.1 研究の背景と目的

日々放送される映像量の増加につれ，再利用や検索を考慮してそれらを整理して蓄積する必要が高まっている．とりわけ，その内容の速報性と利用価値を考えると，ニュース映像への自動的索引付けを行う価値は高い．

筆者らはニュース映像に対して，映像とそれに附随する言語情報を統合的に利用した自動的索引付け手法の構築を目指している．同様の試みは，既に実用的段階に達しているInformediaプロジェクト [9] のNews-on-Demandシステム [10] に代表されるように活発に行われている．しかしそれらの索引付け手法は，全文検索的であったりワードスポッティングや出現頻度に基づくキーワード抽出によるものであったりし，言語情報主導の色合いが強く，映像内容を考慮したものは少ない．この点を考慮して筆者らは，ショット[1]分類に基づく索引付け手法 [3] を提案してきた．この手法は，明示的に記述した一定の条件に従って，ショットを数通りの映像的に類似した典型的ショットに分類し，それぞれの典型的分類に応じた意味属性をもつ言語情報をキーワードとして附与するものであり，Nakamuraら [11] によっても同様の試みが行われている．しかし，これらの手法では分類条件を天下り的に与えるため，分類数が少なくならざるを得ず，かつ恣意的になりがちであるという問題があった．

そこで現在，あらかじめ教師映像データから映像分類条件，つまり画像的特徴量と映像内容を表す言語情報の意味属性との関係を統計的に学習して対応付けておくことにより，索引付けを行いたい未知の映像の画像的特徴量から映像内容を推測する手がかりとし，適切な意味属性をもつキーワードを附与する機構を検討中である．この機構は，ショット分類に基づく索引付け手法の拡張であるのみならず，言語情報を伴わない映像に対してもおおよその内容を推測することが可能であり，大雑把な索引付けすら可能になることが期待される．

本稿では，このような索引付け機構の概要を紹介し，映像分類の観点から，画像的特徴量と映像内容を示す字幕の語義との関係を統計的に学習する手法の提案と，初期段階の実験結果を報告する．

1.2 関連研究の紹介

画像・映像分類の研究はこれまでも多く行われてきた．しかしそれらの多くは，単に画像的特徴量に基づいたクラスタリングなどを行った結果を分類するものである．本研究では，単なる映像分類にとどまらず，概念分類との対応関係を考慮する点でこれらの手法と一線を画している．

初期の関連研究としては，形容詞を中心とした印象語と画像的特徴量の対応を求める栗田ら [5] による絵画データベースに関する研究があるが，ニュース映像データのように具体的事物（主に名詞）を対象とする場合とは問題点やその解決法が異なる．

さらに，孟らによる手法 [6] のように，事例に基づき統計的に分類条件を学習する類似手法もあるが，教師映像データの分類ラベルは人間が事前に与える必要がある．

また，森らによる手法 [7] は，まず語の共起関係により単語クラスタ空間を作成し，次に単語クラスタ空間中の単語間距離に基づき画像に附随する説明文の類似度を計算し，それを反映した画像的特徴量空間中で説明文が類似した画像がクラスタ化される．このように洗練された手法であるものの，単語ではなく文との関係を見ているため，汎用性に欠ける．また，共起関係に基づき形成された単語クラスタの概念分類としての妥当性が明らかではない．

[1] 画像的に連続な映像の最小単位を指す．画像的に不連続なカットに挟まれる区間である．

2 概念分類と画像的特徴の対応付け

2.1 教師映像データからの対応付けの学習

図 1に概念分類と画像的特徴の対応付けの機構を示す.

まず，対応付けを学習するための教師データとして，映像内容を具体的に描写する言語情報（ここではニュース映像中の字幕を利用）を含むショットを解析する．対象となる各ショットの映像からは画像的特徴量ベクトルを抽出し，字幕からは映像内容の概念的分類を行なう．次に，大量の教師映像から，このような概念分類と画像的特徴量ベクトル群との対を作成し，各概念分類を代表するような特徴量ベクトルを学習する．

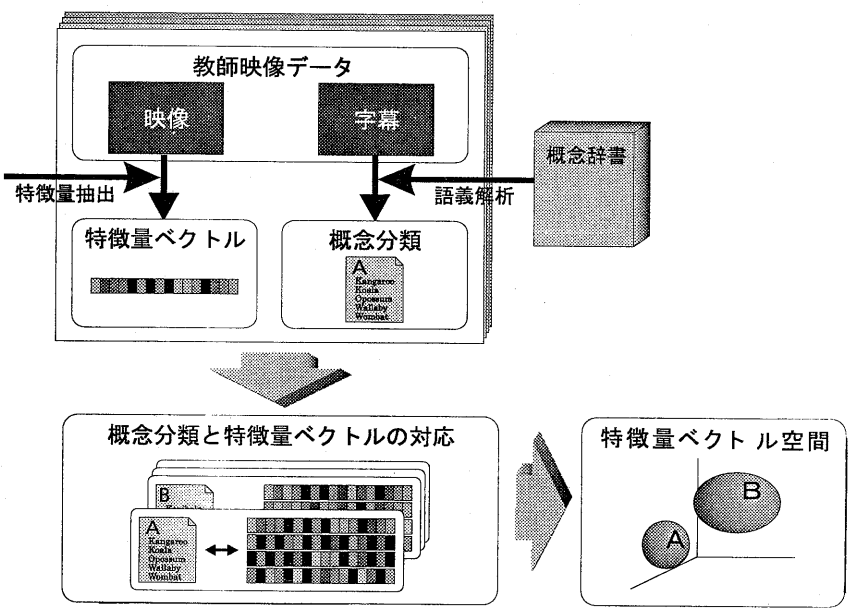

図 1: 概念分類と画像的特徴の対応付け機構

2.2 予備実験

以上のような学習機構の予備実験として，75 分間のニュース映像について，以下の条件の下に字幕の概念分類と画像的特徴の対応を調べた.

- 概念分類には，分類語彙表 [1] の分類項目を使用
- 画像的特徴量として，ショットの先頭フレーム中に存在する比較的大きい顔領域の個数のみを使用
- 概念分類中の語が主に以下のいずれかに属するかに応じて顔の個数との関連を調べることにより，分類性能を評価
  - 人物に関するもの … 1 人の人間を示す
  - 集合に関するもの … 2 人以上の人間の集合を示す
  - その他のもの

表 1: 予備実験の結果：概念分類項目と画像的特徴量の対応（累積度数上位 30%）

顔の個数	分類語彙表中の概念分類項目の見出し		
	人物に関するもの	集合に関するもの	その他のもの
なし	——	「意見，決定，調査，承認」	「地名」「単位」
1個	「長」「人名」	——	「事務所，市場，駅」「地名」
2個	「身分」「人間」	「宣言，報告，噂」	「原理，規則，方法，制度，習慣，計画」「社寺，学校」「家，宿，教室」「地名」「単位」「数」「金銭」
3個以上	——	「意見，決定，調査，承認」「話，議論，批評，説明」「集会，出欠」「議会」「約束，交渉，賛成」	「地名」

表 1に上記の条件にしたがって得られた顔の個数と，対応する概念分類項目のうち累積度数で上位30%に入るものの対応を示す．ここで，分類語彙表の分類項目名を概念分類項目名として表示しているが，一部のものは複数の分類項目を統合して独自に命名してある．また，それぞれの概念分類項目が上記のいずれの分類に属するかは，分類に属する語を基準に著者らの常識に従って主観的に判断した．

この結果を見ると，以下のような特徴が見られる．

- 人物に関する概念分類項目に属する字幕が対応するショットには，1, 2 個の顔が存在する．
- 集合に関する概念分類項目に属する字幕が対応するショットには原則として，2 個以上の顔が存在する．唯一の例外として,「意見，決定，調査，承認」に顔「なし」が対応してしまっているのは，会合などの映像では人々の背後から撮影されることがあり，大勢の人物が存在するにも関わらず，顔領域が検出できなかったためである．
- その他の概念分類項目に属する字幕には場所や組織を示すものが多かったが，これらは顔の個数で分類できる性質のものではない．これらに関しては，他の様々な画像的特徴を利用することにより分類することを目指す．

以上の予備実験により，顔の個数という非常に簡単な特徴のみを用いているものの，比較的良好な対応関係が得られたため，次章ではより多くの画像的特徴を用いた実験を行なう．

3 概念分類と画像的特徴の統計的対応付け実験

3.1 実験手法

本章では，自動抽出した画像的特徴量から概念分類と画像的特徴の対応関係を学習する課程と，学習の結果得られた各概念を代表する画像的特徴量ベクトルとの類似度を計測することによる，評価用画像の概念分類推定の評価実験結果を示す．

3.1.1 用いる画像的特徴量

本実験では，色ヒストグラム（出現頻度分布）と色コリログラム（共起頻度分布）の 2 種類の画像的特徴量を用いる．これらの特徴量は高速に自動抽出可能なものである．

以下の定義中，特徴量を抽出する対象画像を F，縦横のピクセル数を各々m, n，各ピクセルの色の階調を c_{max}とする．また，F中のピクセル p のとる色を $c(p)$ と表す．

表 2: 実験に用いた画像的特徴量

特徴量	解像度（$m \times n$）	c_{max}	d	次元数
色ヒストグラム	320×240	64	——	64
色コリログラム	80×60	16	1, 2, 3, 4	1,024
合計				1,088

- **色ヒストグラムの定義**
  色ヒストグラム $\{H(c_i) \mid i = 1, 2, ..., c_{max}\}$ とは，画像 F 中において，色 c_iのピクセルが存在する確率であり，次式のように定義され，c_{max}次元で表される．

$$H(c_i) \equiv Pr\{(x, y) \mid p \in F, c(p) = c_i\} = \frac{\mid \{p \mid p \in F, c(p) = c_i\} \mid}{mn}$$

- **色コリログラムの定義** [8]
  色コリログラム $\{C_d(c_{i,j}) \mid i, j = 1, 2, ..., c_{max}\}$ とは，画像 F中において，距離 d 離れた 2 つのピクセル p_a, p_b間で色 c_i, c_jが共起する確率であり，次式のように定義され，c_{max}^2次元で表される．

$$\begin{aligned} C_d(c_{i,j}) &\equiv Pr\{(p_a, p_b) \mid p_a, p_b \in F, c(p_a) = c_i, c(p_b) = c_j, \|p_a - p_b\| = d\} \\ &= \frac{\mid \{(p_a, p_b) \mid p_a, p_b \in F, c(p_a) = c_i, c(p_b) = c_j, \|p_a - p_b\| = d\} \mid}{H(c_i)N(d)} \end{aligned}$$

  ここで $N(d)$ は，あるピクセルから距離 d 離れて存在するピクセルの個数である．本実験では簡便のために街区距離を用いているため，$N(d) = 8d$ である．

色ヒストグラムは構造情報をまったく保存しないものの，画像全体の色調を良く表す特徴量である．一方，色コリログラムは近隣の色の共起頻度に基づくため，局所的な構造情報をやや抽象的な方法で保存する特徴量である．一例として，色ヒストグラムでは「日の丸」と「水玉模様」を区別できないのに対し，色コリログラムでは区別できるという特徴がある．本実験では，$(m, n) = (320, 240)$ の大きさの RGB 表色画像から，表 2のような次元で各々の特徴量を抽出した．

3.1.2 字幕による画像の概念分類

本実験では，ニュース映像中に出現する字幕を利用して，そのショットの映像内容の概念分類を行なう．この分類結果が，概念分類と画像的特徴の対応付けの際の分類基準となる．一般に日本語において，名詞句の語義は末尾の名詞によって決定されることが多い [4]．そこで，以下の手順で字幕の解析を行なう．

1. 字幕に対して，日本語形態素解析システム JUMAN [2] による形態素解析を適用
2. 末尾に名詞が存在する字幕に対して，その名詞が概念分類体系中のどの分類項目に属するか検索

このようにしてショット単位に概念分類項目が対応付けられるが，多義語である場合や字幕が複数存在する場合は，複数の概念分類が対応することになる．

概念分類体系として，国立国語研究所編纂の分類語彙表 [1] を用いた．分類語彙表は，36,780 の語が 4 大分類，798 分類項目からなる分類体系であり，1 つの分類段落には 10 程度の概念的に類似した語が含まれている．以下では概念分類項目を $X.Y$（X：大分類，Y：分類項目）のように記し，本実験ではこの単位で概念分類を扱う．なお，分類項目 Yは最大で 4 桁からなる番号であり，画像的特徴との対応付けは，$Y_0, Y_0Y_1, Y_0Y_1Y_2$，そして存在する場合は $Y_0Y_1Y_2Y_3$のそれぞれに対して行なった[2]．

[2]このうち，本来分類語彙表で区別して分類項目名を附与してあるものは，Y_0と $Y_0Y_1Y_2$または $Y_0Y_1Y_2Y_3$のみである．

3.1.3 評価用画像と概念分類項目との距離計算

3.1.1 と **3.1.2** の結果，各概念分類項目に対応するショットの特徴量ベクトル群が得られる．この集合と評価用画像の特徴量ベクトルとの距離を求めることにより，各概念分類との距離を計算する．ベクトル群を代表する値として，正規分布を仮定し，ベクトルの各要素の平均と分散からなる 2 つのベクトルを採用する．ベクトル群を構成する n 個のベクトルを $\boldsymbol{f}(i)(i=1,2,...,n)$ とすると，平均ベクトル $\boldsymbol{\mu}$ と分散ベクトル $\boldsymbol{\sigma^2}$ は次式のように定義され，平均ベクトルによりベクトル群を代表し，分散ベクトルにより分布の広がり具合に応じた距離の正規化を行なう．

$$\boldsymbol{\mu}=\left\{\mu_j \left| \mu_j=\sum_{i=1}^{n}\frac{f_j(i)}{n}, j=(1,2,...,1088)\right.\right\}, \boldsymbol{\sigma^2}=\left\{\sigma_j^2 \left| \sigma_j^2=\sum_{i=1}^{n}\frac{(f_j(i)-\mu_j)^2}{n}, j=(1,2,...,1088)\right.\right\}$$

これらに基づき，次式により標準正規分布に正規化することにより，分布の広がりを考慮して，各概念分類と評価用画像の特徴量ベクトル $\boldsymbol{f}$ の距離 D を求める．

$$D=\sqrt{\sum_{j=1}^{1088}\left((f_j-\mu_j)\exp\left(\frac{1-\sigma_j^2}{2\sigma_j^2}(f_j-\mu_j)^2\right)\right)^2}$$

3.2 実験結果

実験素材として，15 分間のテレビニュース映像 5 本を用意し，各々について以下の手順で実験を行なった．

1. カット検出を行ない，映像をショットに分割．
2. 各ショットの先頭フレームに対して，**3.1.1** で記した要領で 1,088 次元の特徴量ベクトルを抽出．
3. 各ショット内に出現する全字幕に対して，**3.1.2** で記した要領で概念分類を解析．
4. 手順 2. の結果に基づき，手順 1. で得られた各概念分類に属するショットの特徴量ベクトルを集計．
5. 得られた概念分類と画像的特徴量ベクトルとの対応に対して，**3.1.3** に記した要領で評価用画像の特徴量ベクトルからの距離を計算．

評価手法として以下の 2 通りを用意し，表 3に実験結果を示す．

1. 映像 1～5 から標本数 $n \geq 9$ の概念分類の対応関係を集計し，同じ映像集合を用いて評価．
2. 映像 2～5 から標本数 $n \geq 7$ の概念分類の対応関係を集計し，映像 1 を用いて評価．

表 3: 概念分類と画像的特徴対応付け実験の結果

評価手法	評価手法 1.			評価手法 2.		
対応付けに用いたデータ	映像 1～5（212 件）					
評価に用いたデータ	映像 1～5（212 件）			映像 1（32 件）		
分類項目	項目数	1 位率	3 位率	項目数	1 位率	3 位率
(1) $X.Y_0$	3	81%	—	3	63%	—
(2) $X.Y_0Y_1$	10	91%	—	8	19%	—
(3) $X.Y_0Y_1Y_2$	5	78%	—	4	43%	—
(4) $X.Y_0Y_1Y_2Y_3$	2	100%	—	1	100%	—
(5) (1) ～ (4) のいずれか	20	87%	98%	16	53%	79%

表中，項目数とは，定めた最低標本数以上のベクトル群を形成した概念分類の件数であり，1位率，3位率とは各々1位，3位以内に正解が存在した割合である．正解は，対応付けと同様に，**3.1.2** で記した要領で字幕を解析した結果を利用した．

この結果を見ると，以下のような特徴が見られる．

- $X.Y_0Y_1Y_2Y_3$の値は分類項目数が少ないため，この結果からは正解率の善し悪しは判断できない．
- 正解率が高いことが期待される評価手法1.において，$X.Y_0$と $X.Y_0Y_1Y_2$の値が若干低いのは，分散が大きいことや分類間の重なりが大きいことによると思われる．
- 評価方法2.において，概念分類項目が細かくなる（Yの桁数が増える）につれ抽象度が下がるため，正解率が下がることは予想されていたが，$X.Y_0Y_1$の値が極端に低いのは，対応付けに用いたデータが少ないことにも一因があると思われる．

4 結論

本稿では，概念分類と画像的特徴の対応付けを行なう機構の提案と実験結果を示した．実験に用いたデータ数が少ないため，本格的に統計的な処理を行なうのが困難であることや，現時点では特徴量ベクトルの各要素への重み付けを行なっていないことから必ずしも良い結果は得られなかったが，本格的な実験を行なうための基礎データが得られた．今後は，より良い結果を得るべくこれらの点を解決していくとともに，他の様々な特徴量の採用，より高度な対応付け機構の検討を進めていく．

参考文献

[1] 国立国語研究所：国立国語研究所言語処理データ集5：分類語彙表［フロッピーディスク版］，秀英出版 (1993).

[2] 松本裕治，黒橋禎夫，山路　治，妙木　裕，長尾　眞：日本語形態素解析システム JUMAN，ver.3.2 (1997).

[3] Ide, I., Yamamoto, K. and Tanaka, H.: Automatic Video Indexing Based on Shot Classification, *Proc. JSSST First Intl. Conf. on Advanced Multimedia Content Processing*, pp.90–105, (1998).

[4] 井手一郎，田中英彦：末尾の名詞に着目したテレビニュース字幕の語義解析，情報処理学会論文誌，Vol.39, No.8, pp.2543–2546 (1998).

[5] 栗田多喜夫，加藤俊一，福田郁美，板倉あゆみ：印象語による絵画データベースの検索，情報処理学会論文誌，Vol.33, No.11, pp.1373–1383 (1992).

[6] 孟　洋，佐藤真一，坂内正夫：事例を用いた映像シーン分類手法とその評価，第56回情報処理学会全国大会論文集（2），p.176 (1998).

[7] 森　靖英，高橋裕信，岡　隆一：画像と単語の空間配置データベースに基づく画像理解の試み，第4回電子情報通信学会知能情報メディアシンポジウム論文集，pp.127–132 (1998).

[8] Huang, J., Kumar, S. R., Mitra, M., Zhu, W. J. and Zabih, R.: Image Indexing Using Color Correlograms, *Proc. IEEE Computer Vision and Pattern Recognition Conf.*, pp.762–768 (1997).

[9] The Informedia Project, http://www.informedia.cs.cmu.edu/.

[10] Hauptmann, A. G. and Witbrock, M. J.: Informedia News-on-Demand: Using Speech Recognition to Create a Digital Video Library, *Proc. AAAI'97 Spring Symp. on Intelligent Integration and Use of Text, Image, Video and Audio Corpora*, pp.120–126 (1997).

[11] Nakamura, Y. and Kanade, T.: Semantic Analysis for Video Contents Extraction —Spotting by Association in News Video—, *Proc. Fifth ACM Intl. Multimedia Conf.*, pp.393–402 (1997).

[12] Satoh, S., Nakamura, Y. and Kanade, T.: Name-It: Naming and Detecting Faces in Video by the Integration of Image and Natural Language Processing, *Proc. Fifteenth Intl. Joint Conf. on Artificial Intelligence*, pp.1488–1493 (1997).