2007－CVIM－159 (27)
2007／5／15

# Bag-of-keypointsによるTRECVIDデータに対する映像認識

湯 志遠† 柳井啓司†

TRECVIDとは，共通テストコレクションを用いたビデオ検索技術に関する研究開発促進のための国際ワークショップである。本研究では，TRECVIDの中の4つのタスクのうちの1つである，提供されたニュース映像から切り出されたフレーム画像の中から特定の事象と一致する画像を抽出するhigh-level feature extraction task (高次特徴抽出タスク) について扱う。本研究では，このタスクに対して，bag-of-keypointsモデルによる画像認識を適用し，結果の評価を行う。認識する事象はTRECVID2006の評価対象になっているSports, Weather, Officeなど20種類とした。分類手法としては，最近傍法とSVMを用い，両者の結果を比較した。

# Video Recognition For TRECVID By Using Bag-of-keypoints

Zhiyuan Tang† and Keiji Yanai†

TRECVID is an international workshop for research and development promotion about video search technology. TRECVID provides common test data and four kinds of tasks, and participants compete on their results regarding the given tasks. In this research, we tackle the high-level feature extraction task which is one of four tasks in TRECVID. In the high-level feature extraction, we extract frame images in which a specific kind of features appears from the provided news video. For this task, we applied the bag-of-keypoints model which is a state-of-the-art method on generic object recognition, and used nearest neighbor classifier and SVM as classifiers. The features we made experiments on are 20 kinds of features including Sports, Weather, Office, Mountain, Animal and Car.

## 1. はじめに

### 1.1 背景

地上波デジタル放送の開始や，HDDやDVDなどのメディアの普及および次世代の大容量メディアの発売とともに，一般の個人が大量のデジタル映像データを蓄積することができるようになった。

一方，デジタル映像データの大量化とともに，蓄積した大量の映像データの中から，閲覧したい映像を探し出すことが問題となっている。そのために，大量の映像データの中身を自動的に解析して検索を行う映像検索技術への要望が急速に高まっている。この要望に応じて，共通テストコレクションを用いたビデオ検索に関するワークショップTRECVIDが開催されている。

### 1.2 目的

本研究では，TRECVID 2006[1]のhigh-level feature extraction task(高次特徴抽出タスク)を取り上げ，テストコレクションとして提供されたニュース映像から切り出されたフレーム画像の中から特定の事象と一致する画像を抽出するタスクに取り組む．このタスクは，一般的な画像認識[2]の一種と考えることができるため，最新の画像認識手法であるbag-of-keypoints[3]手法を適用して実験を行い，結果の評価を行う。また，比較のために，bag-of-keypointsと同様にヒストグラムに基づくカラーヒストグラム[4]特徴，ニュース映像の音声から音声認識によって自動生成した音声テキストのbag-of-words特徴による実験も行い，結果を比較する。

### 1.3 研究の概要

本研究では，与えられたフレーム画像データの中から特定の事象と一致する画像を抽出するシステムを実現する。データとしては国際映像処理コンテストTRECVIDのデータを用いる。TRECVID2006のデータから学習画像データとテスト画像データを用意し，学習画像データの特徴を用いて，テスト画像データから対応する画像を探し出す。特徴量としてカラーヒストグラムとSIFT(Scale Invariant Feature Transform)[5]特徴で構成したbag-of- keypointsを用い，分類手法として最近傍法とSVM(Support Vector Machine)[6]を使用する。抽出する事象はTRECVID2006の評価対象になっているSports, Weather, Officeなど20種類とした。

† 電気通信大学大学院 電気通信学研究科 情報工学専攻
Department of Computer Science, The University of Electro-Communications

最初に，予備実験としてカラーヒストグラム及び最近傍法を使用したが，十分な結果ではなかった。そこで，SIFT 特徴を抽出し，サイズが 500，1000，1500 のコードブックを構成し，最近傍法と SVM を両方とも使用した。また，比較のために，音声テキストから作成した bag-of-words による単語ベクトル作成し，同様に SVM を用いて実験を行った.

実験の結果，特徴量に関しては，SIFT 特徴の bag-of-keypoints を用いると，カラーヒストグラムとワードベクトルより大幅に良い結果が得られることが分かった。分類手法に関しては，SVM が最近傍法より高い性能を示した。

## 2. TRECVID2006

TRECVID とは，アメリカの国立標準技術研究所 NIST(National Institute of Technology) の研究部門が行うテキスト検索ワークショップ TREC(Text REtrieval Contest) から派生したビデオ映像検索ワークショップである。毎年共通のタスクおよび各タスクに対する評価基準を設定している。

2006 年開催された TRECVID2006 は，2006 年の 8 月中旬が応募締め切りになった。High-level feature extraction タスクには，全 54 グループが参加した。

### 2.1 実験データ

TRECVID では，各タスクの実験を行うための映像データとして，アメリカのニュース番組 CNN や NBC，中国語およびアラビア語のニュース番組を中心として大量の MPEG-1 ファイルが用意されている。さらに，映像データを放送日にしたがって 2 つのグループに大別している。この 2 グループの内，時系列的に古い方 (2004 年秋のニュース映像) を「学習データ」，新しい方 (2005 年秋のニュース映像) を「テストデータ」と定義している。

TRECVID2006 で与えられた映像の本数は，学習データ 137 本，テストデータ 259 本の合計 396 本である。それぞれ，約 80 時間，約 160 時間の膨大な映像データである。映像はショットに分割され，各ショットの代表的な画像が TRECVID から提供される。1 本の映像から 500 枚程度の画像が選出され，学習画像，テスト画像合計 17 万枚以上の画像が与えられる。これらのデータに加え，音声認識の結果のテキスト (中国語やアラビヤ語は英語に翻訳されている) のデータも与えられている (図 1)。本研究では，この中の静止画像データと音声テキストデータを使用する。

**図 1** 例:TRECVID2006 の実験データ (2005.11.24 CNN LIVEFROM)

### 2.2 タ ス ク

TRECVID2006 では，

- shot boundary determination
  実験対象映像ファイルに含まれる各ショット間の切り替え点を自動的に検出するタスク
- high-level feature extraction
  指定された事象が出現する個所を，テスト画像データから検出するタスク
- search
  与えられたクエリを満たすショットを効率的に検索するシステムを開発するタスク
- rushes exploitation
  未編集のラッシュ映像を検索，要約，編集するための支援ツールを開発するタスク

の 4 つのタスクが用意されている。

本研究では，この中の一般的な画像認識タスクである high-level feature extraction(高次特徴抽出タスク) を取り組む。

### 2.3 High-level feature extraction

TRECVID2006 では，14 万枚以上のテスト画像に対して，**Sports**, Entertainment, **Weather**, Court, **Office**, **Meeting**, Studio, Outdoor, Building, **Desert**, Vegetation, **Mountain**, Road, Sky, Snow, Urban, **Waterscape_Waterfront**, Crowd, Face, Person, Government-Leader, **Corporate-Leader**, **Police_Security**, **Military**, Prisoner, **Animal**, **Computer_TV-screen**, **Flag-US**, **Airplane**, **Car**, Bus, **Truck**, Boat_Ship, Walking_Running, **People-Marching**, **Explosion_Fire**, Natural-Disaster, **Maps**, **Charts** の 39 種類の課題すべてについて認識を行わなければならない。

評価基準としては，送付された検索結果 (順位付けされた分類結果の上位 2000 枚の画像) の平均適合率を用いる。平均適合率の算出に必要な正解データは，プーリング方式によって構築されている。プーリング方式とは，参加者が送付した実験結果に含まれるショットのみを評価の対象として，TRECVID の評価者が目で検出対象である事象である事象の出現有無を確認した結果が，正解として採用されるというものである。

TRECVID2006 では 39 種類の内，太字の 20 種類のみを評価対象としている。なお，評価対象の 20 種類

は結果の提出締め切り後に主催者が選んだもので，事前には知られていない。そのため，参加者は 39 種類すべてについて認識を行う必要がある。結果発表後に，評価データが公開される。本研究では，TRECVID2006 において実際に評価が行われたこの 20 種類のみについて取り組むこととする。

## 3. 関連研究

近年，物体認識において，SIFT(Scale Invariant Feature Transform)[5] 特徴量を代表とする局所特徴量表現による研究が盛んに行われている。局所特徴量を用いた物体認識手法は，従来のカラーヒストグラムや領域ベースの手法に比べて，極めて高い性能を示すことが知られている。特に，2004 年に G. Csurka らによって提案された bag-of-words モデル[3] は，簡単で実装が容易なアルゴリズムであるにも関わらず，高い識別性能を示すことが示されている。

SIFT とは，David Lowe によって提案された特徴点とそれに付随する特徴ベクトルの抽出法である。その名が示す通り，画像の拡大縮小，回転や視点のいずれに対して，ロバストであるという特徴を持つ。

G. Csurka らは，局所パターンを SIFT 特徴量で表現し，7 つのクラスの 1776 枚分の画像のすべての特徴量を k-means でクラスタリングして，1000 種類の code book を作成した。分類手法として，SVM と Naïve Bayes を用いた。結果として，複数の同種類物体の画像や遮蔽，部分および方向の物体の画像すべてに対して，精度よく分類した。

また，L. Fei-Fei[7] らは，同じく，局所パターンを SIFT 特徴量で表現し，13 クラスの学習画像 650 枚分の画像のすべての特徴量を k-means クラスタリングして，174 種類の code book を作成し，確率的文書分類手法の LDA を用いて，13 種類のシーンを 64%の精度で分類した。

本研究では，以上の研究を取り入れ，SIFT 特徴を用いて構成した bag-of-keypoints を使用し，分類器として SVM を用いて，TRECVID2006 のデータに対して高次特徴抽出実験を行う。

## 4. 研究の方針

画像認識の一般的な方針として，まず，学習画像と検索対象画像 (テスト画像) を用意し，各画像から特徴を抽出し，学習画像の特徴に基づいて，テスト画像の分類を行う。つまり，コンピュータに学習画像の特徴を用いて，事象を示す画像を勉強させて，テスト画像から事象と一致するかどうかを判断することより，テスト画像を分類する。本研究では，

- 画像のカラーヒストグラム+最近傍法
- 画像の SIFT 特徴の bag-of-keypoints+SVM(Support Vector Machine:SVM)
- 画像の SIFT 特徴の bag-of-keypoints+最近傍法
- 音声テキストデータの bag-of-keypoinnts+SVM

を用いて，実験を構成する。

本研究で使用する実験の主な手法である画像の bag-of-keypoint を用いる手法の流れとしては，図 2 に示している。

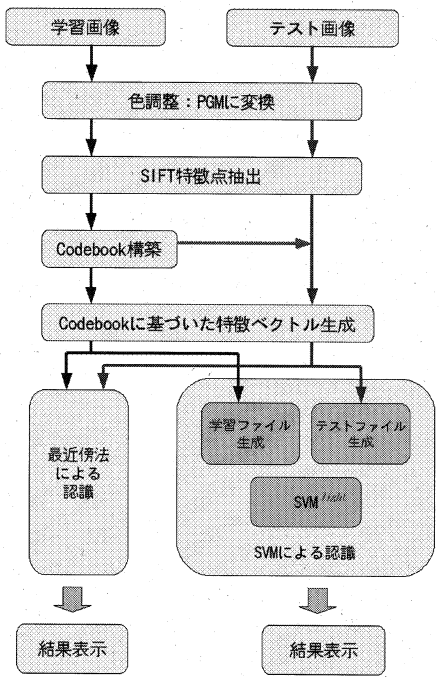


**図 2** システムの流れ

まず，前処理として，すべての学習データとテストデータを jpg 形式から pgm 形式に変換して SIFT 特徴を抽出する。つぎに，学習データの SIFT 特徴を用いて，codebook を構築する。そして，生成された codebook を利用して，各画像の特徴ベクトルを作って，最近傍法や SVM によって認識を行う。

## 5. 画像特徴量

### 5.1 カラーヒストグラム

カラーヒストグラム[4] は色の分布を表示するヒストグラムである。RGB 色空間を用いるとすると。RGB の 3 つのパラメータで色を表現できる。各色の階調を 0 から 255 までとすると，これらの 3 つのパラメータで 3 次元空間を構成すると考えれば，(0,0,0) である黒を原点として，(255,255,255) までの立方体と考えることができる。このとき，RGB が各軸になる。

本研究では，RGB 各軸を 4 分割する。すなわち，

RGB それぞれ，0 から 63 までの間を 0，64 から 127 までの間を 1，128 から 191 の間を 2，192 から 255 までの間を 3 とする。こうすると，$256 \times 256 \times 256$ フルカラー画像を $4 \times 4 \times 4$ に次元を減らす。

そして，参照画像のカラーヒストグラムは次の式

$$M_i'(i = 0 \cdots I-1), I = 4^3$$

とすると，$\sum_i M_i'$ の値はヒストグラムの性質上，その領域の画素数となる。

次に，各ヒストグラムの要素の値を画素数っで割った正規化ヒストグラムを次の式のように作成する：

$$M_i = \frac{M_i'}{\sum_j M_j'}$$

ここで，$\sum_{i=I}^{I} M_i$ は 0,1 に正規化されている。カラー値 0〜63 に対してピクセルの頻度が表示される。

### 5.2 SIFT 特徴

SIFT(Scale Invariant Feature Transform)[5] とは，David Lowe によって提案された特徴点とそれに付随する特徴ベクトルの抽出法である。その名が示す通り，画像の拡大縮小，回転や視点のいずれに対して，ロバストであるという特徴を持つ。

特徴を計算する主なステップは

(1) Scale-space extrema detection
画像の各ピクセルを近隣の 8 ピクセルを比較し，さらに同じシリーズの画像の対応の 9 個のピクセルを比較するによる小塊探知する

(2) keypoint localization
スケール空間の極値から keypoint を選択する

(3) orientation assignment
各 keypoint に関しては，16 × 16 ウィンドウで勾配方向のヒストグラムを計算する

(4) keypoint descriptor
128 次元のベクトルで記述する

の 4 つに分かれている

### 5.3 bag-of-keypoints

カラーヒストグラムは画像全体を表現するに対し，SIFT 特徴は画像の部分の特徴だけを表現する。SIFT 特徴のような局所領域の特徴量のみを用いて，画像をモデル化するために，本研究では bag-of-keypoints[3] の考え方を導入する。

Bag-of-keypoints は，統計的言語処理における bag-of-words[8] のアナロジーである。Bag-of-words で語順を無視して文章を単語の集合と考えるのと同様に，bag-of-keypoints では，位置を無視して画像を局所特徴 (keypoints) の集合として捉える考え方である。実際の処理においては，局所特徴の特徴ベクトルをベクトル量子化することによって，keypoints を word として扱えるようにする。

主なステップは次のように示している：

(1) 画像の局所特徴を探知，記述する

(2) 局所特徴を既定のクラスタ集団に割り当てる (codebook を作成，その要素を codeword と呼ぶ)

(3) それぞれのクラスタに属する特徴の数を数えて bag-of-keypoints，すなわち，codebook パターンのヒストグラムを構成する

(4) Bag of keypoints を特徴ベクトルとして分類方法に適応する

イメージ的に，図 3 のように示している。

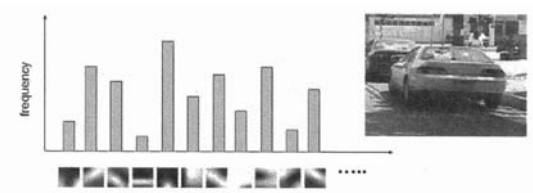


図 3 bag-of-keypoints

#### 5.3.1 k-means 法

Codebook 作成時，局所特徴をクラスタ集団に割り当てるには，k-means 法を用いる。k-means 法とは非階層的なクラスタリング手法の代表例で，アルゴリズムは次のようになる:

(1) 最初に領域数 (k) を決める

(2) 初期クラスタ中心を領域数だけ，ランダムに決める

(3) 各特徴について，クラスタ中心までの距離を計算し，もっとも近いクラスタに配属 (距離として，次に紹介するヒストグラムインタセクションを用いる)

(4) 各クラスタについて，クラスタ中心を計算

(5) 3 4 を帰属が変化しなくなるまで繰り返す

Bag-of-keypoints では，SIFT 特徴のみから Codebook を構築するので，テクスチャに関する情報のみで，色に関する情報が一切含まれていないという特徴がある。

## 6. テキスト特徴量

TRECVID2006 のデータには，一枚のフレーム画像につき，それに対応する一定の時間帯の音声テキストも付いている。認識に用いるために，画像の SIFT 特徴のかわりに，本物の単語を用いて，bag-of-words モデルを構成し，各画像に対応するワードベクトルを求める。

主なステップは次のように示している：

(1) すべての画像のテキストファイル全体について，含まれている単語の頻度を調べる

(2) 頻度が上位 2000 位の単語を取り出し，codebook を作成する (その要素を codeword と呼ぶ)

(3) それぞれの画像テキストファイルから，codeword とマッチする単語の数を数えて bag-of-words，すなわち，codebook パターンのヒストグラムを構成する。

(4) Bag of words を特徴ベクトルとして分類法を適用する

## 7. 分類手法

### 7.1 ヒストグラムインタセクションを用いた最近傍法

#### 7.1.1 ヒストグラインタセクション

1 で正規化した $l$ 個の bin をヒストグラム a と b があり，$i(i=0\cdots l)$ に対して，それぞれの bin が $a_i$，$b_i$ とすると，$a_i$ と $b_i$ の小さい方をとり，これを $i$ が 0 から $l$ まで繰り返し，最後に和を計算する。この総和はヒストグラムインタセクションとなる。

ヒストグラムインタセクションは次のような式で表示できる：

$$S=\sum_{i=1}^{l} min(a_i,b_i)$$

S は 1 と近いほど，a と b の類似度が高い。

#### 7.1.2 最近傍法

最近傍法 (Nearest Neighbor : NN) とは，「特徴空間上で近接しているパターン同士はその性質も互いに似ている」，という仮定に基づいた識別方法である。

n 個のパターンがその所属するクラスとともに $(x_1,\theta_1),(x_2,\theta_2),\cdots,(x_n,\theta n)$ と与えられていたとする。ただし，$\theta_p \in \{\omega_1,\omega_2,\cdots,\omega_c\}(p=1,\cdots,n)$ である。このとき，最近傍法は次式で表される。

$$\min_{p=1,\cdots,n}\{D(x,x_p)\}=D(x,x_k)\Longrightarrow x\in\theta_k \quad (1)$$

$$x_k\in\{x_1,x_2,\cdots,x_n\}$$
$$\theta_k\in\{\theta_1,\theta_2,\cdots,\theta_n\}$$

ここで，$D(x,x_p)$ は $x$ と $x_p$ との距離を表し，式 (1) における $x_k$ は $x$ の最近傍である。

### 7.2 SVM

SVM(Support Vector Machine)[6] は，2 つのクラスを識別する識別器を構成するための学習法であり，1960 年代に Vapnik 等が考案した Optimal Separating Hyperplane を起源とし，1990 年代にカーネル学習法と組み合わせた非線形識別手法へと拡張された。そして，SVM は現在知られている多くのパターン認識手法の中で，最も認識性能の優れた学習モデルの 1 つである。カーネルトリックにより，非線形の識別関数が構成でき，「マージン最大化」を用いることで，未学習サンプルに対しても高い認識性能が得られる。

SVM の学習/分類には，SVM のソフトウェアツールが手軽に利用可能である。本研究では，SVM $^{light}$[9] というソフトウェアツールを用いて，学習を行っている。

## 8. 実　　験

### 8.1 学習データとテストデータ

TRECVID2006 が提供したデータを用いる。データはあらかじめ「学習データ」と「テストデータ」に分けられ，学習データに関して，39 種類の事象の内，TRECVID2006 の評価対象となった 20 種類の事象を使用する。20 の事象それぞれに対して正 (positive) であるか負 (Negative) であるかのリストも渡されている。

シーン切り替えの瞬間に現れる単色の画像を除外すると，実験で使用する学習データは 40991 枚であり，テストデータは 78000 枚である。

### 8.2 評価指標

TRECVID では，一般的に用いられる適合率 (precision) と再現率 (recall) ではなく，1 位から $N$ 位までの適合率の平均値である平均適合率が評価指標として用いられる．平均適合率は，着目する枚数を $N$ とし，1〜$k$ 位までの正解画像の枚数を $OK_k$ とすると，

$$MeanAveragePrecision=\frac{1}{N}\sum_{k=1}^{N}\frac{OK_k}{k} \quad (2)$$
$$(N=2000)$$

と定義される。TRECVID では，N は 2000 が用いられる．ただし，実際には各認識対象に関して正解フレーム画像がそれぞれ 2000 枚以上存在する保証はないため，この平均適合率は絶対的な適合率の平均値というよりも，相対的な比較のための評価指標という意味合いが濃くなっている．

実験結果の評価は，提供されている TRECVID2006 の結果の正解データと評価用のプログラムを用いて行う。認識結果ファイルを規定の形式に変換し，正解データと一緒に引数として評価用プログラムに与えて，簡単に推定平均適合率を求めることができる。

### 8.3 実験結果の表示例

実験の結果はすべてテキストファイル形式となる。これだけでは，内容の確認ができない。結果ファイルに収納された画像を一目瞭然に確認するため，CGI を作り，Web ブラウザで表示する形にする。結果のファイル名を入力して，結果ファイルに類似度でソートされた画像と類似度が順番に表示される。図 4，図 5 及び図 6 は Mountain，Waterscape_Waterfront，Explosion_Fire の表示結果の例である。

### 8.4 結果評価

実験結果について，TRECVID2006 が提供した評価プログラムを用いて，推定平均適合率を求め，TRCVID2006 の結果と比較した。表 1 のように結果を示す。

表 1 において，左から右へ，各列はそれぞれ，評価となる特徴，カラーヒストグラム及び最近傍法による認識結果，codebook サイズが 500 の bag-of-keypoints と

表 1　各手法の結果の infAP と TRECVID2006 の比較

| Features | Nearest Neighbor | | | | SVM | | | | TRECVID 2006 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Bag-of-keypoints | | | Bag-of-keypoints | | | | | |
| | color | 500 | 1000 | 1500 | 500 | 1000 | 1500 | word | median | max |
| 1.Sports | 0.0002 | **0.0091** | 0.0089 | 0.0087 | 0.0236 | 0.0296 | **0.0389** | 0.0179 | 0.254 | 0.495 |
| 3.Weather | 0.0003 | 0.0277 | 0.0458 | **0.0699** | 0.0120 | 0.0147 | **0.0308** | 0.0529 | 0.253 | 0.623 |
| 5.Office | 0.0003 | 0.0000 | 0.0000 | 0.0000 | 0.0011 | 0.0009 | **0.0013** | 0.0011 | 0.004 | 0.153 |
| 6.Meeting | 0.0001 | **0.0035** | 0.0020 | 0.0015 | 0.0320 | 0.0296 | **0.0341** | 0.0049 | 0.111 | 0.314 |
| 10.Desert | 0.0001 | 0.0010 | 0.0004 | 0.0005 | 0.0142 | **0.0246** | 0.0187 | 0.0001 | 0.021 | 0.169 |
| 12.Mountain | 0.0006 | 0.0036 | **0.0050** | 0.0042 | 0.0343 | 0.0359 | **0.0433** | 0.0000 | 0.038 | 0.202 |
| 17.Waterscape | 0.0042 | **0.0002** | 0.0002 | 0.0001 | 0.0699 | 0.0538 | **0.0783** | 0.0002 | 0.039 | 0.275 |
| 23.Police | 0.0013 | **0.0017** | 0.0009 | 0.0006 | 0.0006 | 0.0009 | **0.0010** | 0.0002 | 0.007 | 0.063 |
| 24.Military | 0.0018 | **0.0015** | 0.0013 | 0.0012 | 0.0148 | 0.0176 | **0.0194** | 0.0153 | 0.049 | 0.213 |
| 26.Animal | 0.0001 | 0.0007 | **0.0009** | 0.0006 | 0.0033 | **0.0044** | 0.0041 | 0.0002 | 0.004 | 0.194 |
| 27.Computer | 0.0224 | 0.0011 | 0.0004 | **0.0006** | 0.0092 | 0.0073 | **0.0117** | 0.0012 | 0.114 | 0.320 |
| 28.Flag-US | 0.0006 | **0.0001** | 0.0000 | 0.0000 | 0.0068 | **0.0169** | 0.0068 | 0.0005 | 0.078 | 0.363 |
| 29.Airplane | 0.0002 | 0.0000 | **0.0001** | 0.0000 | 0.0046 | 0.0049 | **0.0067** | 0.0001 | 0.011 | 0.214 |
| 30.Car | 0.0020 | 0.0009 | 0.0008 | **0.0010** | **0.0148** | 0.0100 | 0.0115 | 0.0006 | 0.079 | 0.336 |
| 32.Truck | 0.0017 | **0.0010** | 0.0004 | 0.0005 | **0.0023** | 0.0007 | 0.0004 | 0.0002 | 0.019 | 0.133 |
| 35.People | 0.0000 | **0.0195** | 0.0172 | 0.0174 | 0.0206 | 0.0196 | **0.0247** | 0.0000 | 0.020 | 0.106 |
| 36.Explosion | 0.0000 | **0.0004** | 0.0003 | 0.0003 | 0.0163 | **0.0229** | 0.0227 | 0.0002 | 0.025 | 0.250 |
| 38.Maps | 0.0003 | **0.0031** | 0.0029 | 0.0027 | 0.0144 | **0.0183** | 0.0167 | 0.0045 | 0.170 | 0.463 |
| 39.Charts | 0.0043 | 0.0015 | 0.0037 | **0.0090** | **0.0163** | 0.0134 | 0.0128 | 0.0003 | 0.062 | 0.222 |
| mean | 0.0043 | 0.0061 | 0.0063 | **0.0069** | 0.0156 | 0.0163 | **0.0192** | 0.0050 | 0.070 | 0.192 |

最近傍法による認識結果，codebook サイズが 1000 の bag-of-keypoints と最近傍法による認識結果，codebook サイズが 1500 の bag-of-keypoints と最近傍法による認識結果，codebook サイズが 500 の bag-of-keypoints と SVM による認識結果，codebook サイズが 1000 の bag-of-keypoints と SVM による認識結果，codebook サイズが 1500 の bag-of-keypoints と SVM による認識結果，ワードベクトルと SVM による認識結果，TRECVID2006 の参加グループの認識結果の平均値および TRECVID2006 の参加グループの認識結果の最大値である。また，一番下の行は 20 種類の結果の平均値である。

## 9. 考　　察

比較の結果から見ると，全般的に，特徴量に SIFT 特徴を使用するのはカラーヒストグラムとワードベクトルを使用するより実験結果がよい。また，認識方法に SVM を使用するのは最近傍法を使用するより性能がいいことが明らかである。SVM の認識性能が一般的に優れていることも証明できた。そして，bag-of-keypoints を用いるとき，最近傍法を用いると，codebooksize を 500 にするのは結果がよいが，SVM を用いると codebook のサイズを 1500 にすると，もっともいい結果が得られる。

カラーヒストグラムを特徴量に使用するのはもっとも単純な手法である。色が学習画像と似ている画像しか認識できない。風景以外の認識に向いていないと言われている。結果の中に，課題番号が 27 である Computer_TV-screen について，カラーヒストグラムを用いた手法が他の手法を使うより認識結果がよいという例外がある。その原因はこの事象が色情報に強く依存しているためと考えられる。TRECVID に使用するデータはニュース映像から切り取ったフレーム画像であるため，図 8 のように背景にスクリーンが入っている画像が多く，色が似ている画像が多い。よって，カラーヒストグラムを使用した手法のほうが良い結果が得られた。一方，SIFT 特徴の Bag-of-keypoints を使用した手法は色情報をまったく利用しないために，十分な結果が得られていない (図 9)。

また，Weather についてワードベクトルを用いる認識結果は，ワードベクトルを用いる結果の中，一番良い。Weather の事象に関しては，天気予報の画像が多いため，天気と関連する用語が頻繁に出ることが考えられる。音声テキストデータは個別の事象の認識にとっては重要であることも分かった (図 10)。

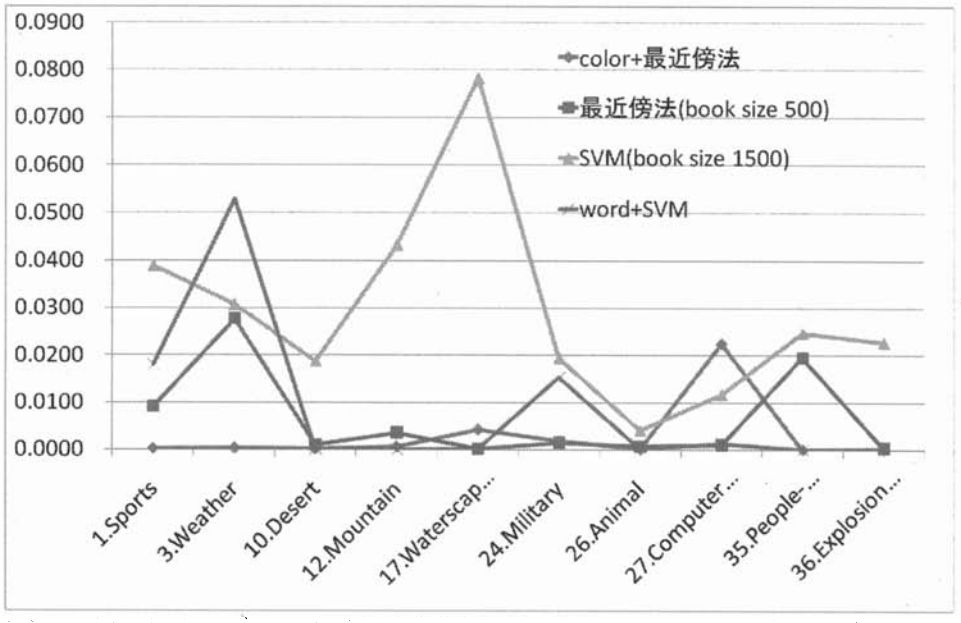


**図 7** 各手法による結果の比較

**図 4** Mountain の認識結果：上から下まで使用した手法は順番に，カラーヒストグラムと最近傍法，bag-of-keypoints と最近傍法，bag-of-keypoints と SVM，ワードベクトルと SVM である。

**図 5** Waterscape_Waterfront：上から下まで使用した手法は順番に，カラーヒストグラムと最近傍法，bag-of-keypoints と最近傍法，bag-of-keypoints と SVM，ワードベクトルと SVM である。

## 10. ま と め

本研究では TRECVID2006 の High-level feature extraction タスクに対し，bag-of-keypoints の手法を適用した。

今回は，特徴量として，ニュース映像から切り出さ

**図 6** Explosion_Fire：上から下まで使用した手法は順番に，カラーヒストグラムと最近傍法，bag-of-keypoints と最近傍法，bag-of-keypoints と SVM，ワードベクトルと SVM である。

**図 8** ComputerTVscreen:カラーヒストグラムと最近傍法による結果

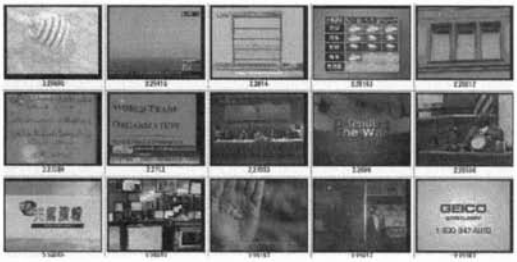

**図 9** ComputerTVscreen: bag-of-keypoints と SVM による結果

**図 10** Weather: ワードベクトルと SVM による結果

れた静止画フレーム画像の画像特徴，及び音声テキストファイルのワードベクトルをそれぞれ独立に利用した。今後は，以上の特徴量の統合，及び動き情報の利用を試みる予定である。

## 参 考 文 献


1) TREC Video Retrieval Evaluation: http://www-nlpir.nist.gov/projects/trecvid/.
2) 柳井啓司：一般物体認識の現状と今後，情報処理学会研究報告，2006-CVIM-155 (2006).
3) Csurka, G., Bray, C., Dance, C. and Fan, L.: Visual categorization with bags of keypoints, *Workshop on Statistical Learning in Computer Vision, ECCV*, pp.1–22 (2004).
4) Swain, M. and Ballard, D.: Color indexing, *International Journal of Computer Vision*, Vol.7, No.1, pp.11–32 (1991).
5) Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, Vol.60, No.2, pp.91–110 (2004).
6) Cristianini, N. and Shawe-Taylor, J.: サポートベクターマシン入門，共立出版 (2005).
7) Fei-Fei, L. and Perona, P.: A Bayesian hierarchical model for learning natural scene categories, *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol.2 (2005).
8) Manning, C. and Schütze, H.: *Foundations of Statistical Natural Language Processing*, The MIT Press (1999).
9) Thorsten Joachims : "SVM $^{light}$ " : http://svmlight.joachims.org/.