TRECVID 高次特徴抽出タスクにおける多種類特徴統合手法の比較

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，色特徴，顔特徴，動き特徴，テキスト特徴および局所特徴による SVM 出力をブースティング，AP weighted fusion などのアルゴリズムで統合する概念認識フレームワークを構成し，TRECVID2008 の概念認識タスクである「高次特徴抽出タスク」に適用し，Airplane，Boat Ship，Classroom などの 20 種類の概念の認識を行った．また，フレームワークの通常の SVM の代りに，各入力特徴を重み付き線形カーネルの和によって統合する Multiple Kernel Learning SVM を導入し，直接統合，認識を行った．実験では，これらの手法による認識結果を比較した結果，最も良い結果として推定平均適合率 infAP として 0.0801 が得られた．この結果は TRECVID2008 の参加者の 39 チームのうち，第 12 位の結果に相当する．

Comparison of Multi-feature Fusion Methods in the High-level Feature Extraction Task of the TRECVID

Zhiyuan Tang† and Keiji Yanai†

The TRECVID is an international workshop for research and development promotion about video search technology. It provides common test data and four kinds of tasks, and participants compete on their results regarding the given tasks. In this paper, we propose a multi-feature fusion framework which combines the SVM classifiers from a number of features by fusion algorithms such as boosting, AP weighted fusion. We utilized the framework in the TRECVID2008 high-level feature extraction task to recognize twenty concepts including Airplane, Boat Ship and Classroom. We also implemented the Multiple Kernel Learning SVM in stead of the normal SVM to perform the fusion and recognition together. In the experiments, we compared all the implemented methods. The best results among all the results achieved 0.0801 as the inferred average precision (infAP), which corresponds to the 12th place of all 39 participants of the TRECVID2008 high-level feature extraction task.

1. はじめに

1.1 背景

近年，パソコンの処理能力の向上，インターネット配信高速化，HDD や DVD などの記憶メディアの大容量化などは，マルチメディアデータの爆発を導いている．特に，flickr や YouTube などの画像，ビデオ共有サイトの登場や地上波デジタル放送の開始により，一般ユーザでも大量のデジタル画像や動画データを蓄積できるようになっている．

一方，デジタルデータの大量化とともに，蓄積した大量のデータの中から，閲覧したいデータ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，TRECVID[1] などの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が開催されている．

1.2 目的

本研究では，TRECVID 2008[1] の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概念を含む画像を抽出するタスクに取り組む．

本研究では，画像から抽出された複数の特徴による SVM の出力結果を Boosting，AP weighted fusion などのアルゴリズムで統合する概念認識フレームワークを構成し，TRECVID2008 高次特徴抽出タスクに適用し，認識対象の Airplane flying，Boat ship，Classroom などの 20 種類の概念の認識の精度向上を図る．また，複数の特徴カーネルを統合する MKL SVM も導入し，これらの統合手法の比較を行う．

1.3 研究の概要

本研究は，複数特徴統合フレームワークを構成し，TRECVID2008 の 4 つのタスクのうち，与えられた

† 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

キーフレーム画像データの中から特定の概念と一致する画像を抽出するタスクである高次特徴抽出タスクに適用する．TRECVID2008の実験データに対して，色特徴，顔特徴，動き特徴，テキスト特徴および局所特徴を抽出し，各特徴ごとにSVMによって学習，分類を行って，最後にBoostingやAP weighted fusionなどの統合方法で，各特徴単独のSVMの出力結果を統合し，最終結果を得る．実験では，高次特徴抽出課題での認識対象概念であるAirplane_flying，Boat_ship，Classroomなどの20種類について，特徴統合フレームワークによる評価実験を行った．また，一般的なSVMの代りに，各入力特徴量を重み付き線形カーネルの和によって統合するMKL SVM(Multiple Kernel Learning SVM)を導入し，これらの実験結果の比較を行った．

実験の最も良い結果として，推定平均適合率 infAP=0.0801が得られ，最終的にTRECVID2008の参加者の39チームのうち，第12位の結果に相当する結果を得ることができた．

2. TRECVID

TRECVID(TREC Video Retrieval Workshop)とは，米国の国立標準技術研究所(NIST: National Institute of Standards and Technology)と国防総省(U.S. Department of Defense)主催するテキスト検索ワークショップTREC(Text REtrieval Conference)から派生した，映像検索技術に関する研究を促進する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．

2.1 提供データ

TRECVID2008では，実験を行うための映像データとして，オランダ語のドキュメンタリ，教育番組，科学番組を中心とした大量のMPEG-1ファイルが用意されている．データは学習データとテストデータから2つに分割されており，放送された日時の時系列に従って，古い方を学習データ，新しい方をテストデータとしている．学習データは構築したシステムの学習データとして利用されるため，正解データ(groundtruth data)のメタデータが付随している．一方，テストデータは，TRECVIDワークショップのタスクを実際に行うための実験データである．

2008年開催されたTRECVID2008で与えられた映像の合計時間は約200時間である．これらの映像はさらに約36000個ショットに分割されており，各ショットから切り出されたキーフレームおよび各ショットに対応する音声認識テキストも提供されている．

2.2 タスク

TRECVID2008では，

- 監視イベント検出(Surveillance event detection)
  大量のイベントセットから特定のイベントを検出するタスク
- 高次特徴抽出(High-level feature extraction)
  指定された概念が出現するショットを，テスト画像データから検出するタスク
- 検索(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
- ラッシュ要約(Rushes summarization)
  映像中の重要な部分のサマライズするタスク
- 内容ベースコピー検出(Content-based copy detection)
  同じ映像のコピーを内容ベースで検出するタスク

の5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 高次特徴抽出タスク

TRECVID2008の高次特徴抽出タスクは，参加者に配布された映像データを構成する大量のショットのフレーム画像から，その意味内容に応じて「飛行機」,「船」,「橋」などの概念を含む画像を検出するタスクである．映像意味解析の高度化を目的としたタスクである．対象概念に分類されたショットが学習用映像セットとして与えられ，参加者はこれを用いて適切な検出手法の構築を行う．後ほど，評価用映像セットに対し構築した手法を適用し，クラスに分類した結果(順位付けされた分類結果の上位2000枚のショット)をプロジェクトオーガナイザーに送付することにより，分類結果の精度の相互比較が行われる．

TRECVID2008の高次特徴抽出タスクの認識対象は，Classroom，Bridge，Emergency_Vehicle，Dog，Kitchen，Airplane_flying，Two people，Bus，Driver，Cityscape，Harbor，Telephone，Street，Demonstration_Or_Protest，Hand，Mountain，Nighttime，Boat_Ship，Flower，Singingの20種類がある．参加者は参加者はこの20種類の概念すべてについて認識を行わなければならない．

3. 関連研究

TRECVIDにおいては，IBM研究所のMilindら[2]はTRECVID2004ワークショップのサーベイ論文で，ビデオ検索や分類などの分野に大きな改善を与えることを示した．それ以来，複数特徴統合の研究は近年盛んで行われている．2005~2007年3年連続で高次特徴抽出タスクの一位を取った中国の清華大学のWangら[3]は色，テクスチャ，エッジ，SIFTなど合計30個以上の特徴を用いて，それぞれSVMで分類を行い，ブースティングで統合する手法を利用した．アムステルダム大学のSnoekら[4]は色情報が付いているSIFTおよび特徴点抽出方法の異なった4種類のSFITを特徴量として用いて，特徴量のヒストグラムを統合して得られた16000次元の巨大ベクトルで，SVMを使って実験をして，TRECVID2008の一位を取った．インテル研究所のLiら[5]は，カラーヒストグラム，エッ

ジヒストグラム，HoG ヒストグラムなどのヒストグラムパターン特徴にマルコフ連鎖モデルを適用することによって，TRECVID2005 の実験データで実験を行い，それぞれの単独の特徴での認識性能を向上させることができることを示した．しかし，統合を行わなかったため，コンテストの順位については十分な結果が得られなかった．また，台湾国際大学の Weng ら[6] は，TRECVID2006 のデータを用いて，Columbia374 の概念アノテーションを導入し，決定木のグラフィカルモデルを解くことで，文脈と時間情報の統合を行った．結果は 30%上がったが，異なったデータを利用したので，TRECVID ワークショップに typeC としてしか参加できない (TRECVID が提供するデータのみを使用するのは typeA である)．

また，TRECVID 以外では，Varma ら[7] は MKL (Multiple Kernel Learning) の学習方法を提案した．彼らは，Oxford flower や Caltech101 などの実験データに対して，MKL を利用し複数の特徴の最適重みを計算し，特徴統合の実験を行い，顕著な性能向上を果たした．

本研究は，Wang ら[3] と Li ら Snoek ら[4] の研究を取り入れ，TRECVID2008 の実験データに対して，色特徴，顔特徴，動き特徴，テキスト特徴および局所特徴を抽出し，SVM によって，モデリングし，最後に AdaBoost などの 3 つ統合方法で，SVM の分類子を統合し，認識結果を得る．また，各入力特徴量をカーネルの重み付き線形結合によって統合する，Varma ら[7] の手法を取り込んだ SVM である MKL SVM(Multiple KernelLearning SVM) を導入し，SVM の内部で統合を実験を行う．

4. 多種類特徴統合フレームワーク

4.1 フレームワークの概要

フレームワークの処理の流れは図 1 に示す．まず，主催者から提供された学習データとテストデータのキーフレーム画像に対して，色特徴，テクスチャ特徴，顔特徴，局所特徴及びテキスト特徴を抽出する．また，ショットの再分割を行い，キーフレーム以外，その前後 1 秒間のフレームも取得し，動き特徴を抽出し，全部で 6 種類のカテゴリの特徴を使用する．

次に，抽出された特徴ベクトルをそれぞれ単独で SVM(Support Vector Machine) に代入し，モデリングを行う．

最後に，特徴の数分の SVM の出力 (SVM 分類子) を本研究で提案した Adaboost のシンプルバージョンやオリジナル Adaboost 及び AP weighted fusion などの統合アルゴリズムを用いて統合し，認識結果を求める．

また，もう 1 つの統合手法として，図 2 の通常の SVM 出力の統合フレームワークの代りに，各入力特徴を重み付き線形カーネルの和によって統合する MKL SVM(Multiple Kernel Learning SVM) をフレームワークに導入し，直接統合，認識を行い，結果を求める．

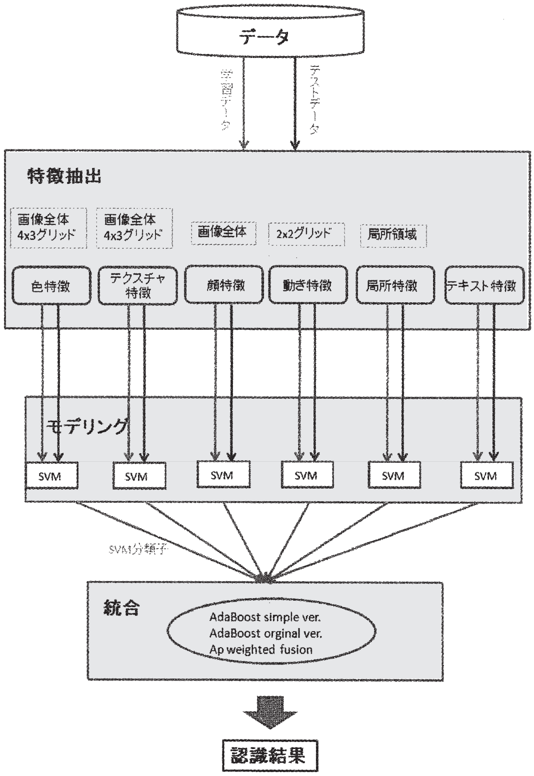

図 1 SVM 分類子の統合によるフレームワーク

4.2 画像特徴量

- **色特徴**：色特徴としては，(1) キーフレーム画像の RGB 色空間の各軸を 4 分割し，各領域の画素値の頻度を表現する 64 次元の正規化カラーヒストグラム[8] (2)RGB のそれぞれの次元を独立に平均，分散，3 次モーメントおよび 4 次モーメントを計算，連結することによって生成された 12 次元のカラーモーメントの 2 種類を使用する．
- **局所パターン特徴**：局所パターン特徴としては，SIFT 特徴[9] を利用する．特徴点抽出アルゴリズムによって，特徴点を抽出し，その周囲のパターンをコード化する．特徴点に関して，本研究は図 3 のように，DOG，ランダムとグリッドの 3 つの角度から検出を行う．局所パターンを用いて画像をモデル化するために，bag-of-keypoints 手法を導入する．本研究では，SIFT 特徴の生成には，SIFT++[10] というアプリケーションを利用する．特徴点の位置情報および PGM 形式の白黒画像を引数として実行すると，本実験で使用する 3 種類の SIFT 特徴を全部生成することが可能である．
- **テキスト特徴**：TRECVID2008 が提供した音声

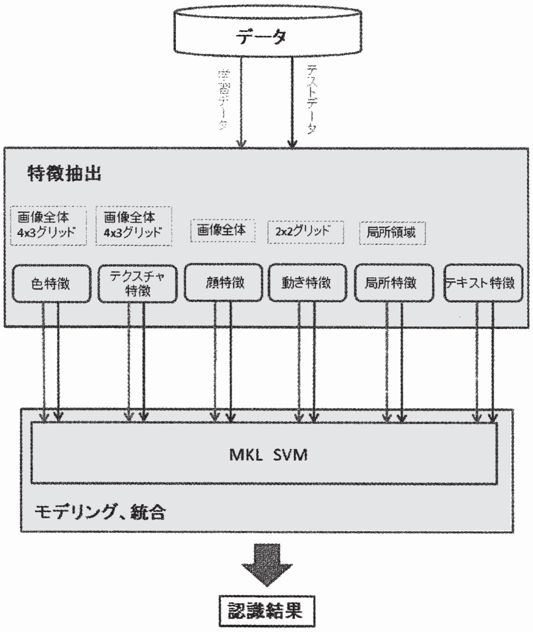

図 2 MKL SVM によるフレームワーク

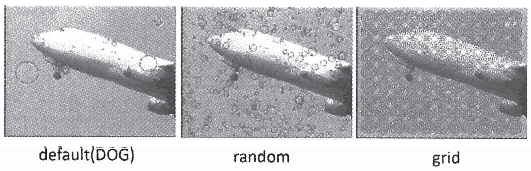

図 3 特徴点の種類

テキストデータに対して，tf-idf アルゴリズムを適用することによって得られたワードベクトルを使用する．

- **動き特徴**：キーフレーム前後 0.5 秒のオプティカルフローを求め，角度空間を 12 分割し，動きの大きさで投票して得られたヒストグラムを利用する．本研究では，Lucas-Kanade 法[11] を用いてオプティカルフローを抽出する．
- **顔特徴**：キーフレーム画像から抽出された顔の数を使用する．本研究では，顔の検出は OpenCV(インテルが開発・公開しているオープンソースのコンピュータビジョン向けライブラリ)[12] の顔画像検出プログラムルーチンを利用する．
- **テクスチャ特徴**：6 方向，4 周期のガボールフィルタを用いて生成された 24 次元のガボールベクトルを使用する．

4.3 統合手法

フレームワークの統合段階において，ブースティングと AP weighted fusion の統合手法を用いる．また，特徴を統合するには，通常の SVM の代りに，MKL SVM を導入する．本節では，それぞれの統合手法について詳細を説明する．

4.3.1 ブースティング

ブースティングは，統計的学習理論の分野において，多数の精度の低い識別器を組み合わせて非常に精度の高い識別器を作るアンサンプル学習方式であり，1990 年に Schapire により初めて導入された[13]．その後，Freund による改良[14] を経て,1995 年に Freund と Schapire によって AdaBoost(Adaptive Boosting)[15] と呼ばれるアルゴリズムを提案された．Adaboost は現実の識別問題に適用され，高い識別性能が示されている．本研究では，SVM を弱識別機として用いる AdaBoost のオリジナルバージョンに加えて，AdaBoost の学習段階の resampling 部分を簡約化したシンプルバージョンを導入し，SVM の分類子の統合に応用する．

AdaBoost(オリジナルバージョン) 本来の AdaBoost は分類木のような重み付き学習データから学習可能な弱識別機を用いるが，本研究では学習データを重みに応じてサンプリングしてそれを各ラウンドの学習データと見なし，SVM を弱識別機として用いる．本研究に適用するオリジナルの AdaBoost のアルゴリズムは図 4 に示す．学習データの特徴ベクトルに対し，正解データなら「1」，不正解データなら「-1」のラベルを付ける．それぞれのデータは合計が 1 になるような正規化された初期重みを持つ．AdaBoost の各ラウンドで，誤認識率から学習データの重みを決定し，さらに，これらのデータからランダムサンプリングされたデータを用いて，すべての特徴量についてそれぞれ別々に再学習を行い，最も分類性能の高い弱識別器を見つける．最後に，各ラウンドで選択された弱識別器を用いて新しい弱識別器を生成する．本論文では，以後 AdaBoost のオリジナルバージョンを「originalAbaBoost」と呼ぶことにする．

AdaBoost(シンプルバージョン) AdaBoost のオリジナルバージョンでは，ループのラウンド毎に各特徴量について SVM の再学習をするので，計算コストが大きいという問題があった．そこで，本研究では，オリジナルの AdaBoost の再学習の部分を省略し，resampling の代わりに，reweighting 方式に改造した簡単化した AdaBoost についても実験を行う．具体的な違いは，AdaBoost のループを入る前に，すべての学習データを用いて学習を行っておくことである．各ラウンドにおいて，最も分類性能の高い弱識別器を見つけ，エラー評価時にのみ用いて学習データの重みを更新する．ループ内で SVM の学習を行わないため，AdaBoost の「シンプルバージョン」と呼ぶ．アルゴリズムは図 5 に示す．本論文では，以後 AdaBoost のシンプルバージョンを「simpleAbaBoost」と呼ぶことにする．

4.3.2 AP weighted fusion

フレームワークに，AdaBoost 以外に，AP wighted fusion[16] も用いて統合を行う．AP weighted fusion

入力：学習データセット X ($|X| = n$), その中，positive データセットを X_1，negative データを X_0 とする．ここで $y(x_i) = 1$ if $x_i \in X_1$ and $y(x_i) = -1$ if $x_i \in X_0$.

1. 重み w を初期化：

$$w_1(x_i) = \begin{cases} \frac{1}{2|X_1|} & if\ x_i \in X_1 \\ \frac{1}{2|X_0|} & if\ x_i \in X_0 \end{cases} \quad (1)$$

2. For t = 1, ...,T:
   - 重みを正規化：

$$w_{t,i} = \frac{w_{t,i}}{\sum_{m=1}^{n} w_{t,m}} \quad (2)$$

   - X から重み $w_{t,i}$ に応じてデータ集合 X' を選択し，各特徴量 j に対し，分類器 h_j を学習する
   - エラーを評価する:

$$e_i = |h_j(x_i) - y_i|/2 \quad (3)$$
$$\epsilon_j = \sum_i w_i e_i \quad (4)$$

   - 最もエラーが小さい h_t を選択する
   - 各学習データを更新する：

$$\beta_t = \frac{\epsilon_t}{1-\epsilon_t} \quad (5)$$
$$w_{t+1,i} = w_{t,i}\beta_t^{1-e_i} \quad (6)$$

3. 出力：

テストデータ x_k に対し：

$$h(x_k) = \sum_{t=1}^{T} \alpha_t h_t(x_k) - \frac{1}{2}\sum_{t=1}^{T}\alpha_t \quad (7)$$
$$\text{where } \alpha_t = log\frac{1}{\beta_t}$$

図 4 AdaBoost のオリジナルバージョンのアルゴリズム

入力：学習データセット X ($|X| = n$), その中，positive データセットを X_1，negative データを X_0 とする．ここで $y(x_i) = 1$ if $x_i \in X_1$ and $y(x_i) = -1$ if $x_i \in X_0$.

1. 重み w を初期化する：

$$w_1(x_i) = \begin{cases} \frac{1}{2|X_1|} & if\ x_i \in X_1 \\ \frac{1}{2|X_0|} & if\ x_i \in X_0 \end{cases} \quad (1)$$

2. 各特徴量 j について，X 全体を用いて分類器 h_j を学習する
3. For t = 1, ...,T:
   - 重みを正規化する：

$$w_{t,i} = \frac{w_{t,i}}{\sum_{m=1}^{n} w_{t,m}} \quad (2)$$

   - エラーを評価する:

$$e_i = |h_j(x_i) - y_i|/2 \quad (3)$$
$$\epsilon_j = \sum_i w_i e_i \quad (4)$$

   - 最もエラーが小さい分類器 h_t を選択する．
   - 各学習データの重みを更新する：

$$\beta_t = \frac{\epsilon_t}{1-\epsilon_t} \quad (5)$$
$$w_{t+1,i} = w_{t,i}\beta_t^{1-e_i} \quad (6)$$

4. 出力:

テストデータ x_k に対し：

$$h(x_k) = \sum_{t=1}^{T} \alpha_t h_t(x_k) - \frac{1}{2}\sum_{t=1}^{T}\alpha_t \quad (7)$$
$$\text{where } \alpha_t = log\frac{1}{\beta_t}$$

図 5 AdaBoost のシンプルバージョンのアルゴリズム

とは，名前が示す通り，各特徴の AP(Average Precision：平均適合率) を重みとする統合アルゴリズムである．手順としては，次のようになる．

(1) 学習データを用いて，各単独特徴に対して，学習行い，AP を計算する．
(2) 計算された AP を合計が 1 になるように正規化する．
(3) 正規化された AP を重みとして，各特徴の SVM 分類子に付け，統合を行う．

4.3.3 Multiple Kernel Learning

Multiple Kernel Learning (MKL) は複数の SVM カーネルを線形結合によって統合するときに，最適な重みを学習によって求める手法である．複数の SVM カーネル (サブカーネル) を線形結合することより，統合カーネルを生成する式は以下の様になる．

$$K_{combined}(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{K} \beta_j k_j(\mathbf{x}, \mathbf{x}') \quad (1)$$
$$where \beta_j \leq 0, \ \sum_{j=1}^{K} \beta_j = 1$$

学習データが与えられた時に各サブカーネルの重み β_j をどのような値に設定すれば，統合カーネルが最適でになるかというのが MKL 問題で，SVM を反復することによって，SVM 学習の最適化の枠組みで同時に最適な β_j を求める方法が提案されている[17]．

MKL では，各カーネルに与えられる特徴ベクトルは同じでも，カーネル毎に異なっていてもどちらでも構わないので，Varma ら[7] は画像認識における特徴統合に MKL を利用することを提案している．．本研究では，それと同様に MKL を多種類特徴の統合手法として利用することとする．

5. 実 験

5.1 学習データとテストデータ

本研究の実験では，TRECVID2008 が提供した映像データをショット分割によって生成されたショットから抽出されたフレーム画像およびショットと対応する音声認識テキストを用いる．フレームの画像の数として，学習データは約 33,000 枚，テストデータは約 36,000 枚が含まれている．また，分類学習のため，認識対象の 20 種類の概念に対して正 (positive) であるか負 (negative) であるかのリストも学習データと一緒に提供されている．

5.2 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，

$$Average\ Precision(AP) = \frac{1}{N}\sum_{k=1}^{N}\frac{OK_k}{k} \quad (2)$$

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．実際には，TRECVID2008 では，サンプリングによって一部の分類結果のみを評価することによって求めた推定平均適合率 (Inferred Average Precision：infAP) を用いて評価を行う．

実験結果の評価は，提供されている TRECVID2008 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．

5.3 特徴抽出

実験では，前述の 6 種類のカテゴリの画像特徴を異なった角度，領域から抽出した．抽出された特徴はそれぞれの簡略表示と説明は表 1 に示す．

表 1 特徴の簡略表示及び説明

簡略表示	説明
ch_globle	画像全体領域から抽出されたカラーヒストグラム
ch_grid	画像を 4 × 3 グリッド分割して，抽出されたカラーヒストグラム
cm_globle	画像全体領域から抽出されたカラーモーメント
cm_grid	画像を 4 × 3 グリッド分割して，抽出されたカラーモーメント
face	顔ベクトル
motion	動きベクトル
text	ワードベクトル
BoK_D_D	画像の DoG 点 SIFT 特徴を用いて，DoG 点 SIFT の codebook ベースで構築された BoK モデル
BoK_D_r	画像の DoG 点 SIFT 特徴を用いて，random 点 SIFT の codebook ベースで構築された BoK モデル
BoK_g_g	画像の grid 点 SIFT 特徴を用いて，grid 点 SIFT の codebook ベースで構築された BoK モデル
BoK_r_D	画像の random 点 SIFT 特徴を用いて，DoG 点 SIFT の codebook ベースで構築された BoK モデル
BoK_r_r	画像の random 点 SIFT 特徴を用いて，random 点 SIFT の codebook ベースで構築された BoK モデル
gb_globle	画像全体領域から抽出されたガボール特徴
gb_grid	画像を 4 × 3 グリッド分割して，抽出されたガボール特徴

5.4 実験設定

実験では，

(1) 単独特徴による認識

(2) 4 種類特徴:ch_bloble, ch_grid, face, BoK_D_D の統合

(3) 10 種類特徴:ch_bloble, ch_grid, face, motion, text, Bok_D_D, Bok_D_r, Bok_g_g, Bok_r_D, Bok_r_r の統合

(4) 全種類の特徴の統合

を行う．SVM の実装は SVM^{light}[18] によって実現し，カーネルを χ^2 カーネルに設定する．

5.5 実験結果

各特徴単独で認識した結果 (20 種類の infAP の平均) は表 2 に示す．また，4 種類の統合方法で 4 種類特徴，10 種類特徴および全特徴に対する統合結果は表 3 に示す．また，表 4 と表 5 はそれぞれ Two people と Nighttime に対して，4 種類の統合方法が 4 種類特徴に重み付けの例を示している．

表 2 単独特徴の分類結果

特徴	infAP	特徴	infAP
ch_globle	0.0160	Bok_D_D	0.0157
ch_grid	0.0220	Bok_D_r	0.0129
cm_globle	0.0114	Bok_g_g	0.0125
cm_grid	0.0275	Bok_r_D	0.0108
face	0.0092	Bok_r_r	0.0097
motion	0.0009	gb_globle	0.0154
text	0.0018	gb_grid	0.0342

表 3 統合結果の比較

統合手法	4 種類	10 種類	全特徴
Ada(smp)	**0.0326**	0.0314	0.0319
Ada(org)	0.0345	0.0382	**0.0500**
Apw	0.0495	0.0645	**0.0801**
MKL	0.0441	0.0528 (motion,text を除く)	—

表 4 重み選択の例：Two people

統合手法	ch_globle	ch_grid	face	BoK_D_D	infAP
Ada(smp)	0	0.05240	0.4332	0.0428	0.0164
Ada(org)	0	0	1.0000	0	0.0201
Apw	0.2504	0.2677	0.2542	0.2267	0.0289
MKL	1.0000	0	0	0	0.0007

表 5 重み選択の例：Nighttime

統合手法	ch_globle	ch_grid	face	BoK_D_D	infAP
Ada(smp)	0	0.9864	0	0.0146	0.1264
Ada(org)	0	0.4658	0	0.5342	0.1316
Apw	0.2694	0.2699	0.2005	0.2603	0.1349
MKL	0	0.6854	0	0.3146	0.1184

表 2 より，グリッド分割することによって，位置情報が付いている特徴の性能が分かった．また，動き特徴，テキスト特徴はあまり効果がなかったことも分かった．動き特徴について，特徴抽出する際に，0.5 秒間隔で非常に大きい変動がある画像は省略するように設定したので，全体の実験データから動き特徴を半分ほどしか抽出できなかった．それに，カメラの動きを考慮入れなかったことやノイズなどの原因で，分類は正確にできなかった．テキスト特徴に関して，もとの実験データの半分しかテキストデータが付いていなかった上，それぞれのテキストデータは機械翻訳でオランダ語から英語に翻訳されたもので，画像を表現することは難しいと考えられる．

表 3 より，AdaBoost のシンプルバージョン以外の

統合方法は特徴の種類の増加とともに結果が良くなることが分かった．4 種類の統合方法のうち，最も結果が良かったのは AP weighted fusion となっている．

図 6 と図 7 はそれぞれ Flower と Boat_Ship に対して，統合性能が最も良かった AP weighted fusion による認識結果を示す．

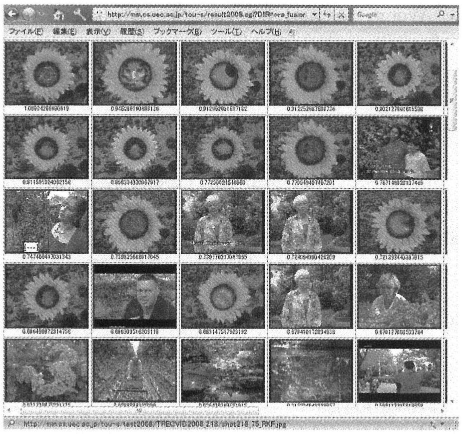

図 6 Flower の認識結果

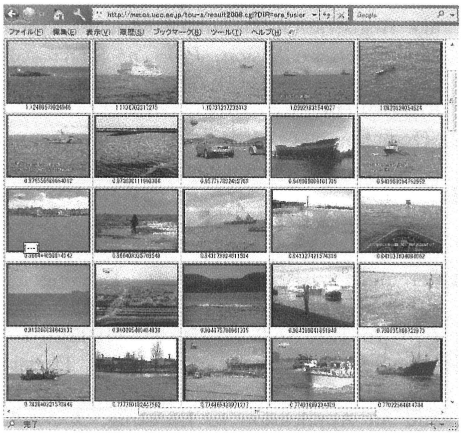

図 7 Boat_Ship

表 6 の各行は 20 種類の対象に対して，AdaBoost の改良版，AdaBoost のオリジナルバージョン，AP weighted fusion，MKL SVM による統合結果および TRECVID2008 の参加チームの結果 infAP の平均値，最高値を示している．一番下の行は各列の平均を示している．本実験の最高結果は，AP weighted fusion の結果の infAP=0.0801 となり，これは全世界 39 チームのうち第 12 位，日本の 8 チームのうち，第 2 位と相当する結果である．

表 6 結果の比較

concept\fusion	smpAda	orgAda	APw	MKL	median	max
01.Classroom	0.0038	0.0015	0.0218	0.0239	0.008	0.152
02.Bridge	0.0055	0.0123	0.0249	0.0175	0.004	0.117
03.E_Vehicle	0.0017	0.0001	0.0062	0.0015	0.003	0.065
04.Dog	0.0188	0.0145	0.1503	0.1192	0.067	0.271
05.Kitchen	0.0053	0.0161	0.0523	0.0389	0.010	0.165
06.Airplane_fly	0.0301	0.0161	0.0255	0.0181	0.029	0.278
07.Two people	0.0385	0.0201	0.0495	0.0007	0.050	0.174
08.Bus	0.0005	0.0007	0.0034	0.0032	0.004	0.119
09.Driver	0.0232	0.0268	0.0731	0.0682	0.046	0.324
10.Cityscape	0.0544	0.0803	0.1292	0.1138	0.059	0.258
11.Harbor	0.0085	0.0080	0.0110	0.0155	0.007	0.182
12.Telephone	0.0022	0.0023	0.0360	0.0168	0.011	0.136
13.Street	0.0760	0.0808	0.1746	0.0001	0.112	0.413
14.Demonstr	0.0126	0.0206	0.0502	0.0746	0.013	0.233
15.Hand	0.0665	0.0779	0.2035	0.0012	0.092	0.377
16.Mountain	0.0354	0.0401	0.0751	0.1154	0.042	0.246
17.Nighttime	0.1004	0.1358	0.1511	0.1571	0.105	0.323
18.Boat_Ship	0.1125	0.1017	0.1655	0.1330	0.093	0.394
19.Flower	0.0887	0.0912	0.1116	0.1154	0.058	0.161
20.Singing	0.0052	0.0168	0.0873	0.0211	0.013	0.258
mean	0.0345	0.0382	0.0801	0.0528	0.043	0.233

5.6 統合性能の比較

実験では，simpleAdaBoost，originalAdaBoost，AP weighted fusion の 3 つのアルゴリズム及び MKL SVM を使用して，複数の特徴の統合実験を行った．まず，AdaBoost アルゴリズムに関して，originalAdaBoost では，毎回サンプリングによって異なる学習データで弱識別器が学習され，ループ毎に選択される弱識別器が変化するために多様な識別器を統合でき，よい学習ができる．一方，simpleAdaBoost では，最初に一度だけ全部の特徴を用いて学習する．その後，エラー率評価時の学習データの重みしか変化しないので，ループ毎に同じ弱識別器が選択されることが起こり，あまり効果良く学習できない．実際の実装では，学習データの数が少なく，AdaBoost に使用する SVM の分類性能が高いので，両方とも任意に 1 つの特徴を選択し，AdaBoost のループから打ち切る傾向があった (例えば，表 4 の AdaBoost オリジナルバージョンの結果)．AP weight fusion でも，同様に，データの数と SVM 性能の理由で，特徴同士では求めた AP は大きな差がなく，重みはほぼ均一に割り当てられてしまった．

4 つの方法の重み選択は表 4，表 5 を通して比較できる．両方の表において，2 種類のブースティングの結果では，ch_globle 特徴は全然選択されていない．原因として考えられるのは，simpleAdaBoost と originalAdaBoost において，データの重みを求める際に，エラー変数や SVM の出力値には曖昧さが存在するので，特徴選択には影響を与えてしまうことである．通常の AdaBoost では，図 4 の式 (3) と図 5 の式 (3) において，e_i は誤分類されたデータの数を表すので，整数になっている．本研究では，条件を緩め，e_i を [0,1] の実数に設定している．しかし，SVM 出力値 $h_j(x_i)$ は実数になっているので，どのように $h_j(x_i)$ を [0,1] の範囲内に変換するかに関しては曖昧さが存在してい

る．一方，表 4 と表 5 において，AP weighted fusion に関して，各概念に対し，それぞれの特徴は 0.2 と 0.3 の間の実数重みを持っている．また，MKL SVM の内部では，最適化問題を解くので，結果は最も良いと予想したが，実際には，表 4 の結果を見ると，Two people のような概念には，face 特徴の重みは 0 となっているので，重みの選択は最適とはいえない．4 つの方法には，どの方法は最適重みを選択できると絶対的な判断はできないが，結果から見ると，AP weighted fusion の infAP は最も大きいので，本実験では最も有効であることが分かった．

6. ま と め

本研究では多種類特徴を統合する概念認識フレームワークを構成し，TRECVID2008 の高次特徴抽出タスクに適用した．そして，統合結果を用いて，フレームワークに使用した 4 種類の統合手法を比較した．その中，AP weighted fusion の結果は infAP=0.0801 まで上がった．最終的に，TRECVID2008 の参加者の 39 チームのうち，第 12 位の結果と相当する結果を得ることができた．

本実験の結果より，テキスト特徴や動き特徴などのような非静止画像特徴は，TRECVID のデータには性能が良くないことが分かった．今後は非静止画像特徴の代りに，静止画像特徴を中心で増やし，統合することにより，より優れた結果を得ることは一つの課題である．また，実験では，MKL SVM を実装したが，理想な結果が得られなかった．MKL SVM の改良やパメータの調整などで，性能向上を図ることはもう一つの課題である．

参 考 文 献

1) TREC Video Retrieval Evaluation: `http://www-nlpir.nist.gov/projects/trecvid/`.
2) Freund, Y., Shapire, R., Gilad-Bachrach, R., Navot, A., Tishby, N., Holland, J., Henning, M., Vinoski, S., Lenaerts, T., Manderick, B. et al.: Experiments with a new boosting algorithm, *Symbolic Visual Learning*, Vol. 10, pp. 81–116 (2002).
3) Wang, D., Liu, X., Luo, L., Li, J. and Zhang, B.: Video diver: generic video indexing with diverse features, *Proc. of ACM MM WS on Multimedia Information Retrieval*, pp. 61–70 (2007).
4) Snoek, C., vande Sande, K., de Rooij, O., Huurnink, B., van Gemert, J., Uijlings, J., He, J., Li, X., Everts, I., Nedovic, V., van Liempt, M., van Balen, R., Yan, F., Tahir, M., Mikolajczyk, K., Kittler, J., de Rijke, M., Geusebroek, J., Gevers, T., Worring, M., Smeulders, A. and Koelma, D.: The MediaMill TRECVID2008 Semantic Video Search Engine, Technical report, ISLA,University of Amsterdam, Amsterdam, The Netherlands (2008).
5) Fei-Fei, L. and Perona, P.: One Step Beyond Histogram: Image Representation by Markov Stationary Features, *Proc. of IEEE Computer Vision and Pattern Recognition* (2008).
6) Weng, M. and Chuang, Y.: Multi-cue fusion for semantic video indexing, *Proc. of the 16th ACM international conference on Multimedia*, pp.71–80 (2008).
7) Varma, M. and Ray, D.: Learning The Discriminative Power-Invariance Trade-Off, *Proc. of IEEE International Conference on Computer Vision*, pp.1–8 (2008).
8) Swain, M. and Ballard, D.: Color indexing, *International Journal of Computer Vision*, Vol.7, No.1, pp.11–32 (1991).
9) Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, Vol.60, No.2, pp.91–110 (2004).
10) A lightweight C++ implementation of SIFT: `http://vision.ucla.edu/~vedaldi/code/siftpp/siftpp.html`.
11) Lucas, B. and Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision, *Proc. of International Joint Conference on Artificial Intelligence*, pp. 674–679 (1981).
12) Open Source Computer Vision Library: `http://www.intel.com/technology/computing/opencv/`.
13) Schapire, R.: The Strength of Weak Learnability, *Machine Learning*, Vol.5, No.2, pp.197–227 (1990).
14) Freund, Y.: Boosting a weak learning algorithm by majority, *Information and Computation*, Vol.121, No.2, pp.256–285 (1995).
15) Freund, Y. and Schapire, R.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, *Journal of Computer and System Sciences*, Vol.55, No.1, pp.119–139 (1997).
16) Wang, M. and Hua, X.: Study on the combination of video concept detectors, *Proc. of the 16th ACM international conference on Multimedia*, pp.647–650 (2008).
17) Sonnenburg, S., Rätsch, G., Schäfer, C. and Schölkopf, B.: Large Scale Multiple Kernel Learning, *Journal of Machine Learning Research*, Vol.7, pp.1531–1565 (2006).
18) Thorsten Joachims : "SVM ^{light}": `http://svmlight.joachims.org/`.