Vol.2011-CVIM-177 No.28
2011/5/19

マルチフレーム認識を用いた動画像認識の分析

樋爪和也[†1]　柳井啓司[†1]

本研究は，近年に提案されたマルチフレーム手法のショット認識に対する有効性の検証が目的である．TRECVID2010 の実験データから SURF，色特徴，時空間特徴を抽出し，それらを Bag-of-Features 表現に変換する．この際，1 枚のキーフレームのみから特徴を抽出を行う従来手法とは異なり，マルチフレーム処理として動画から複数のフレームを取得して特徴を抽出し，統合することで一つの特徴ベクトルを生成した．実験では TRECVID のタスクで指定されている五つの概念を対象とし，キーフレームのみの認識との比較を行った．さらにフレーム取得枚数や選択方法を変化させたときの認識率の変化も検証した．

実験の結果，キーフレームのみの認識に比べて SVM で学習，分類をした場合は認識率が最大で 700%上昇し，MKL-SVM で学習，分類をした場合は最大で 883%上昇した．TRECVID2010 の全チームの平均値と比較した結果，MKL-SVM で学習，分類をした場合に全クラスで値を上回った．

Analysis of video data recognition using multi-frame

Kazuya Hidume[†1] and Keiji Yanai[†1]

In this study, we aim to verify the effectiveness of a multi-frame method for shot recognition proposed in recent years. In the experiments, we extract SURF, color and spatio- temporal features from the TRECVID 2010 video data, and convert them the Bag-of-Features(BoF) representation. In the multi-frame method unlike the conventional method to extract features from only one keyframe, features are extracted from multiple frames which are selected from the video, and one BoF feature vector is generated by integrating these features. In the experiment, we use five kinds of concepts out of 130 TRECVID2010 target concepts and analyze recognition performance in various settings in terms of the number of frames selected from one shot. As a result, compared to the conventional method, the recognition accu- racy in classifying by SVM raised 700% at most and by MKL-SVM advance 883%. The result of MKL-SVM in all class outperformed the average of all the teams in TRECVID2010.

1. はじめに

近年，広く普及している動画サイトでは，主に動画周辺のタグや文章を使用したテキストベースによる映像検索方法が用いられている．しかし，それらの文章はユーザの主観によって付与されるため，部分的に単語を含む場合や意味概念と異なる単語の使い方をしている場合は他のユーザのニーズに対応できない可能性が発生する．そこで WEB サイトに存在する動画に含まれる映像・音響情報を利用する映像検索の研究が盛んに行われている．映像が持つ映像・音響情報に対しパターン認識の技術を活用することで，コンテンツベースの検索を行うことが可能となる．特に TRECVID[1)] と呼ばれる動画認識における国際的な評価ワークショップが存在し，映像検索技術に関する進歩と促進が狙われている．この TRECVID にて成果を挙げている技術の一つとしてマルチフレーム認識がある．

本研究は大量の Web 動画に対して，映っている特定の物体，風景，動作を認識および分類することが目的である．ショット認識において，従来手法では単一フレームのみを扱うのに対し，本研究では複数のフレームを扱うマルチフレーム認識を行うことで，一般物体認識の精度向上を図る．また，フレームの選択枚数および選択方法を変化させることで，マルチフレーム認識の有効性を検証する．なお，本研究は TRECVID で行われるタスクのルールに則って行い，他チームの評価との比較を行うことで効果を検証する．

2. TRECVID

TRECVID（TREC Video Retrieval Evaluation）とは，米国国立標準技術研究所（NIST）と Disruptive Technology Office(DTO) が主催する映像検索技術に関連する研究を促進する競争型ワークショップである．毎年，各国の研究グループが参加し，TRECVID が設定した共通のタスクおよび評価基準に対して各々の手法で研究に取り組む．共通のタスクに対して優劣を競争させ，結果の比較検討および成果の共有を行うことで，映像検索技術の研究を促進させることが目的である．

2010 年に開催された TRECVID2010 では 6 つのタスクが用意され，本研究ではその内の Semantic indexing（SIN）タスクのルールに従う．SIN タスクでは全テストデータの中から「バス」,「教室」,「歌っている人」などの特定の概念を含んだショットを検出するタス

†1 電気通信大学大学院　情報理工学研究科　総合情報学専攻
Department of Informatics, Graduate School of Informatics and Engineering, The University of Electro-Communications

© 2011 Information Processing Society of Japan

Vol.2011-CVIM-177 No.28
2011/5/19

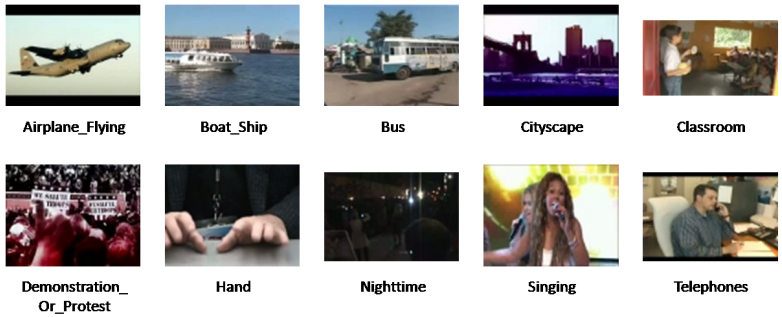

図 1 TRECVID2010 Semantic indexing の認識対象例

クである．TRECVID2010 では認識対象となる概念が 130 種類まで増加された．図 1 はその概念の一例である．また，インターネットアーカイブ映像が MPEG-4/H.264 ファイル形式で学習データ，テストデータ共にそれぞれ 200 時間ずつ配布された．

3. 関 連 研 究

本研究で扱うマルチフレーム認識が TRECVID において使われるようになったのは 2008 年からである．アムステルダム大学の研究チーム[2] はショットのキーフレーム周辺から最大 4 枚の追加フレームを取得し，それらのフレームから SIFT をはじめとする 7 つの特徴量を抽出した．2009 年では取得フレームを 10 枚に増やし，視覚特徴だけでなく音響特徴を新たに追加し，さらに Multiple Kernel Learning 手法の一つである SR-KDA(Spectral Regression combined with Kernel Discriminant Analysis)[3] と MK-FDA(Multiple Kernel Fisher Discriminant Analysis)[4] も採用している．アムステルダム大学はいずれの TRECVID タスクにおいても最高の成績を収めている．これに対し，東京工業大学[5] は 2009 年にショットの全てのフレームから SIFT を抽出している．音響特徴としてメル周波数ケプストラム係数 (MFCC) も抽出し，それぞれをガウス混合分布を使用してモデル化している．この手法により 4 位の結果を残した．

4. 提案手法概要

4.1 マルチフレーム認識

マルチフレーム認識の処理について，図 2 に示す．従来の大量のビデオ映像に対する主

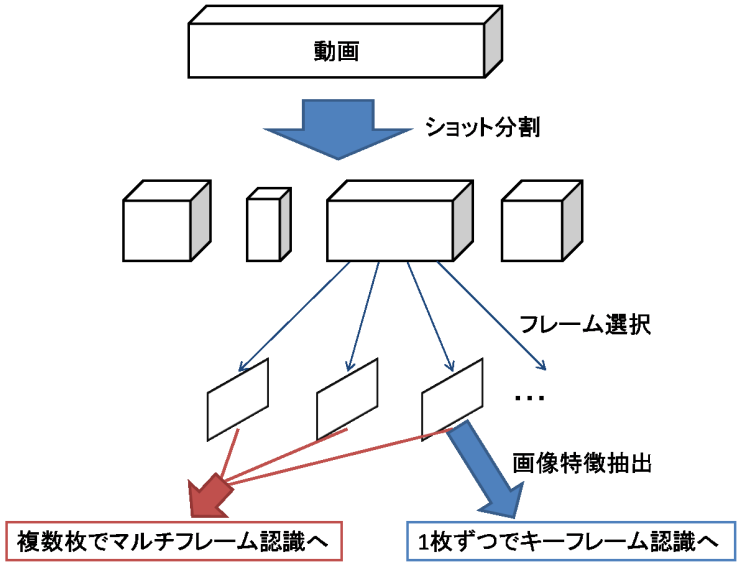

図 2 マルチフレーム処理

な認識手法は，ショット分割を行った後に各ショットの中から何らかの基準によって 1 枚だけ抽出したキーフレームと呼ばれるフレーム画像を使って特徴量を抽出し，認識を行ってきた．しかし，ショットは複数枚のフレームから構成されるため，2 枚以上のフレーム画像から特徴量を抽出することができる．

そこで動画を短時間のシーンごとに分割したショットを作成し，各ショットから複数のフレームを選択する．この数枚のフレームから得た特徴量を統合して一つの特徴ベクトルを作成する．これにより，同一の特徴量だけでも特徴ベクトルの次元を増やすことができる．ここでマルチフレームの選択例を図 3 に示す． 図のように同一ショットから抽出したフレーム画像は，全体的に似通っているが細かい点で差異が生じているのがわかる．このように 1 つのショットから認識対象について複数のアングルの画像特徴を検出できれば，より良い認識結果になると予想される．

本研究では，フレームの選択基準として以下の二つを採用する．

- 1 つのショットから等間隔に N 枚を選択
- 1 つのショットから M フレーム置きに選択

© 2011 Information Processing Society of Japan

Vol.2011-CVIM-177 No.28
2011/5/19


マルチフレーム例 Airplane_Flying


マルチフレーム例 Bus


マルチフレーム例 Hand

図 3 マルチフレーム選択例

等間隔に選択する方法は，ショット全体を $N+1$ 分割した際の境界を選択する．M フレーム置きに選択する方法は，初めに選択したフレームから M フレーム間隔で存在するフレームを選択していく．このときマルチフレームで統合される枚数は，それぞれ N 枚，(ショットのフレーム数 $-1)/M+1$ 枚となる．よって N の値が大きくなるほど，M の値が小さくなるほど統合するフレーム枚数は多くなる．比較に使用するパラメータはそれぞれ $N=3,5,10, M=30,15,10$ である．M の数値はデータセットの動画が 30 フレーム/秒であることに由来する．また実験では，比較として全フレームから特徴量を抽出する方法でも認識を行う．

4.2 処理の流れ

本研究の処理の概要図を図 4 に示す．

まず，ショット中の選択したフレームから SURF および色特徴を，ショット全体から時空間特徴を抽出する．これらを Bag-of-Features 表現に変換するが，SURF および色特徴にはマルチフレームを適用し，複数枚のフレームの特徴量を使用する．この二つの特徴量は空間ピラミッドマッチングを行えるように画像を 2×2 の領域に分割した上で量子化する．時空間特徴は元々マルチフレームから抽出する特徴なので，同様にショット単位での Bag-of-Features ベクトルを生成する．

それぞれの特徴量を用いて SVM または MKL-SVM で学習を行い，各特徴量を統合する

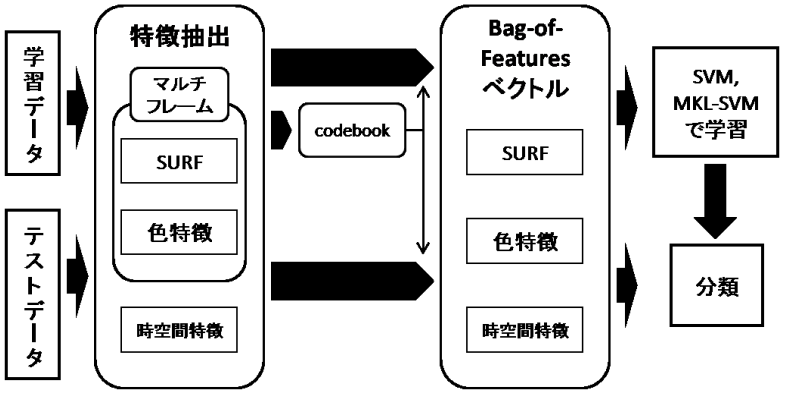

図 4 処理の流れ

際の最適な重みを計算する．SVM のカーネルには性能の高い RBF-χ^2 カーネルを使用する．テスト動画のショットの認識を行い，得られた出力値をそのショットの評価値としてランキング付けを行う．このランキングの上位 2000 ショットを各概念の検出結果とし，適合率を求める．

5. 認識手法

5.1 Bag-of-Features

Bag-of-Features とは，画像を局所特徴の集合として考える手法であり，言語処理における Bag-of-Words の考え方を画像処理に適用したものである．Bag-of-Words は文章中の単語の語順を無視し，単語の出現頻度によって文章を表現する．Bag-of-Features では同様に，画像の局所特徴の位置を無視し，出現頻度をヒストグラム化することで画像を表現する．本研究ではショットに対してマルチフレーム処理を行うため，これを画像単位からショット単位に拡張して扱う．

まず，静画像に対する Bag-of-Features 表現への変換は以下の流れで行う．

(1) 学習画像から局所特徴を抽出する．
(2) すべての局所特徴を k 個にクラスタリングする．この k 個のクラスタ中心の特徴ベクトルを visual words と呼び，visual words の集合を codebook と呼ぶ．
(3) 画像毎に検出した局所特徴をベクトル量子化する．一つ一つの局所特徴に対し，各 visual words の中で最も近い特徴ベクトルに投票を行うことで出現回数のヒストグ

© 2011 Information Processing Society of Japan

ラムを作成する．

(4) 画像の局所特徴の総数でヒストグラムの各 bin を割り，正規化を行う．

本研究で扱うデータはショット単位となっている．マルチフレームの場合，一つのショットから複数枚のフレーム画像を選択し，特徴量を抽出する．ここで得られた全ての画像特徴 (本研究では SURF，色特徴) は一つのヒストグラムに変換される．すなわち，一つのショットから一つの Bag-of-Features ベクトルが作成される．これにより，大量に取得した特徴量を比較的次元数の少ない特徴ベクトルへと変換することができる．また，ショット認識においてはフレーム画像から得られる画像特徴の他に，オーディオ特徴や後述の時空間特徴など静画像には無い特徴量を得ることができる．これらについてもショット単位の Bag-of-Features ベクトルを作成することができる．本研究で扱う特徴量のうち，時空間特徴については学習用ショットから得られた特徴量を使って前述の手順と同様に codebook を作成，ヒストグラムへの投票を行う．

5.2 色 特 徴

色特徴は画像の色分布を表す特徴であり，画像認識においては基本的な特徴である．

本研究では RGB 色空間のカラーヒストグラムを作成する．画像の RGB 値で構成される 3 つのパラメータを Bag-of-Features における局所特徴とみなし，画像の全画素について visual words への投票を行う．

5.3 SURF

SURF(Speeded-Up Robust Feature)[6] とは，特徴点とその周辺の局所特徴を記述した特徴量である．画像の照明変化，スケール変化，回転に対して頑健であり，64 次元の特徴ベクトルで表現される．また，サブ領域を使用することで 128 次元に拡張することもできる．同様の特徴量として SIFT[7] が存在する．SIFT も 128 次元の特徴量であるが，SURF は SIFT に比べて処理が高速であり，また SIFT と同程度の精度がある．

本研究では SURF の特徴点検出は行わず，特徴点およびそのスケールをランダムに 5000 個決定する．これは TRECVID2010 データが Web 動画を扱っているために解像度の低い動画も含まれており，そのような動画からは特徴点がほぼ検出できないという可能性があるためである．

5.4 時空間特徴

時空間特徴は動作認識を行う上で欠かせない特徴である．以下では柳井研究室の野口[8] が提案した時空間特徴抽出手法について説明する．この手法は大量の Web 動画分類に適しており，既存手法に比べて計算コストが低く抑えられている．

検出は次の手順で行われる．

step1 カメラモーション検出
step2 SURF を各フレーム画像から抽出
step3 フレーム間で動きのなかった SURF 点を削除
step4 Delaunay 三角分割
step5 動き特徴抽出
step6 視覚・動き特徴の統合

まず，動画に含まれるズームやパンなどのカメラモーションの検出を行う．Web 動画には手振れのような意図しないカメラモーションが多く含まれており，さらに解像度が低い．これを解決するために，カメラモーションを検出した場合，その特徴を破棄する．検出手法には Lucas-Kanade アルゴリズム[9] を用いる．

次にフレーム画像から SURF を抽出する．この SURF の点の内，時空間特徴として適しているのはフレーム間で動きのある点のみであるため，動きのなかった点は削除する．そして残った点について Dlauney 三角分割を行う．Dlauney 三角分割とは空間内の点を連結して三角形のグループを作成することで，点の特徴だけでなくその周辺の特徴も考慮する手法である．この 3 点の SURF 記述子を視覚特徴とするため，$64 \times 3 = 192$ 次元で表現される．

動き特徴は，まず SURF を抽出したフレームから N フレーム先までのフレームを取得する．上記の SURF 点の動き情報はこの内の $N/2$ フレームから計算される．取得した N フレームを M 分割し，その区間内で Lucas-Kanade アルゴリズムによって特徴点のオプティカルフローを計算する．各区間の動き特徴はオプティカルフローの x,y 成分の正方向および負方向に動きなしを加えた 5 次元で表現される．設定は区間 $N = 5$，分割数 $M = 5$ としており，三角形の面積変化は 5 次元で表現されるため，動き特徴は $20 \times 3 + 5 = 65$ 次元となる．また，動き特徴を回転に対して有効にするため，SURF 記述子で得た回転角を利用してオプティカルフローの回転も行う．

最後に視覚特徴，動き特徴を単純に結合することによって $192 + 65 = 257$ 次元の一つの時空間特徴とする．

5.5 空間ピラミッドマッチング

Bag-of-Features 表現は物体認識において高い性能を示しているが，その性質上，特徴の位置関係を無視してしまう．そこで Lazebnik ら[10] によって空間ピラミッドマッチング手法が提案された．局所特徴の位置を考慮できるように画像をグリッド領域に分割し，各領域内で Bag-of-Features ベクトルを作成する．

© 2011 Information Processing Society of Japan

本研究では動画全体の特徴量として抽出される時空間特徴を除いた，SURF と色特徴に適用することができる．

6. 分 類 器

本研究では，分類器として SVM(Support Vector Machine) と MKL-SVM(Multiple Kernel Learning SVM) を使用する．

6.1 Support Vector Machine

SVM(Support Vector Machine) は 2 値分類問題を解くために考えられた学習アルゴリズムであり，基本的には線型識別器である．現在知られている数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ひとつである．

SVM は線型識別器であるが，特徴ベクトルを変換するカーネルトリックによって非線型への拡張が可能である．特徴ベクトル x を非線型の写像 $\phi(x)$ によって線型識別可能な空間に写像し，その空間で識別を行うという方法をとる．本研究では以下の RBF-χ^2 カーネルを用いる．

$$K(x,y) = exp\left(-\frac{1}{2\sigma^2}\sum_i \frac{||x_i - y_i||^2}{x_i + y_i}\right) \tag{1}$$

6.2 Multiple Kernel Learning

MKL(Multiple Kernel Learning) は複数の SVM カーネル (サブカーネル) を統合することにより，新しい最適なカーネルを求める手法である．新しいカーネルは以下の式で求められる．

$$K_{combined}(x,x') = \sum_{j=1}^{K} \beta_j k_j(x,x') \quad \beta_j \geqq 0, \sum_{j=1}^{K} \beta_j = 1 \tag{2}$$

ここで最適なカーネルを生成するために，サブカーネルに対する重み β_j を学習する必要がある．この問題は MKL 問題と呼ばれている．MKL 問題は β_j のすべての組み合わせを実際に計算することで解くことができる．しかし，特徴やカーネルの数が増えるにつれて β_j の組合わせの数も膨大になってしまうため，現実的な方法とは言いづらい．そこで近年，MKL 問題を凸面最適化問題として解く方法が提案されている．その一つに Sonnenburg[11] らは単一カーネルの SVM 学習を反復することによって，最適な重み β_j を求める手法を提案している．

表 1 データセットの内訳

クラス	ポジティブショット数	ネガティブショット数
Airplane Flying	66	1000
Boat Ship	172	2500
Bus	31	1000
Cityscape	558	5000
Classroom	139	2500

7. 実 験

7.1 データセット

データセットは TRECVID が提供する TRECVID2010 の学習データ，テストデータを利用する．実験にはこのうち，5 個の概念を使用して学習と分類を行う．クラスのカテゴリと，学習に使用するクラス毎のショット数を表 7.1 に示す．ポジティブショットとは各クラスの概念が出現するショットを，ネガティブショットとは出現しないショットを指す．

また，テストデータは全 144,988 ショットである．

7.2 評 価 指 標

TRECVID2006 より，評価指標には推定平均適合率 (Inferred Average Precision : infAP) が用いられている．適合率とは認識されたもののうち正しい割合であり，(クラスに分類された正解データ数)/(分類されたデータ数) で求められる．上位 N 位までの平均適合率は，第 k 位までの適合率 $Precision(k)$ を用いて，

$$infAP = \frac{1}{N}\sum_{k=1}^{N} Precision(k) \tag{3}$$

として求められる．推定平均適合率はさらに，(分類された正解データ数)/(全正解データ数) で求められる分類率を考慮したものであり，全正解データの数が不明である場合にも有効である．なお，TRECVID2010 では評価用のプログラムが配布されている．本実験は TRECVID2010 Semantic Indexing タスクに則り，上位 2000 ショットを評価対象とする．

7.3 実験の設定

本実験で扱う特徴量は SURF，色特徴，時空間特徴の 3 つである．いずれも Bag-of-Features 表現により出現頻度の特徴ベクトルを作成する．本研究では codebook，すなわち Bag-of-Features ベクトルを 500 次元と設定する．このうち，時空間特徴の Bag-of-Features ベクトルはショット全体から抽出するため，フレームの取得枚数の変化にはよらずショット毎に

© 2011 Information Processing Society of Japan

表 2 M フレーム置きに抽出したときの平均フレーム選択枚数，値はポジティブショット/ネガティブショット

	M フレーム置き		
クラス	M=30	M=15	M=10
Airplane_Flying	4.06/5.45	7.62/10.44	11.15/15.45
Boat_Ship	5.40/6.05	10.60/11.72	15.82/17.46
Bus	19.06/5.23	37.65/10.08	56.55/14.94
Cityscape	4.80/7.77	9.20/15.23	13.61/22.67
Classroom	14.11/7.29	27.95/14.23	41.94/21.21
テストデータ	4.99	9.48	14.01

固定の値である．

空間ピラミッドマッチングによる領域のグリッド分割数は 2×2 とする．各領域において 500 次元の特徴ベクトルを作成するため，$500 \times 4 = 2000$ 次元の特徴ベクトルとなる．これを SURF，色特徴に適用する．よって，本実験で扱う特徴ベクトルは $2000 \times 2 + 500 = 4500$ 次元となる．この次元数は以下のすべての実験で共通である．

また，いずれの実験においても SVM，MKL-SVM には RBF-χ^2 カーネルを使用する．

7.3.1 マルチフレームの設定

本研究では，以下の 3 つの実験を行う．学習データ，テストデータ共に同じマルチフレーム抽出方法で特徴抽出を行う．

(1) マルチフレーム認識とキーフレームのみの認識の比較
(2) フレーム選択方法の比較
(3) MKL-SVM による評価

従来手法である 1 枚のキーフレームのみの認識との比較を行い，マルチフレーム認識の有効性を検証する．また，3 章で述べたフレームの二つの選択基準の認識結果の違いを確認する．加えて，一つのショットから全フレームを抽出した場合の結果とも比較を行う．これらの実験はすべて SVM による学習，分類を行う．また，M フレーム置きに抽出する方法について，各クラス毎の平均フレーム選択枚数を表 2 に示す． 対して，MKL-SVM でも同様の設定で実験を行い，フレーム枚数を変化させたときの認識結果と，MKL で学習する特徴量毎の重みの変化を確認する．比較として，TRECVID2010 の全チームの推定平均適合率の平均値 (median) と最大値 (max) を記載する．

7.4 実験結果

SVM で実行した実験結果の推定平均適合率を表 3 に示す．図 5 は表 3 のグラフである．簡略のために従来手法であるキーフレームのみの認識を $N = 1$ と表記する．

表 3 フレーム選択方法の比較

	N 枚抽出				M フレーム置きに抽出			
クラス	N=1	N=3	N=5	N=10	M=30	M=15	M=10	全フレーム抽出
Airplane_Flying	0.0076	0.0263	0.0282	0.0301	0.0331	0.0608	0.0400	0.0414
Boat_Ship	0.0294	0.0334	0.0282	0.0278	0.0220	0.0270	0.0276	0.0274
Bus	0.0021	0.0028	0.0030	0.0032	0.0034	0.0034	0.0043	0.0043
Cityscape	0.0586	0.0818	0.0746	0.0466	0.0785	0.0796	0.0722	0.0889
Classroom	0.0002	0.0006	0.0003	0.0003	0.0012	0.0014	0.0004	0.0006

表 4 MKL-SVM による評価

	N 枚抽出				M フレーム置きに抽出				TRECVID2010	
クラス	N=1	N=3	N=5	N=10	M=30	M=15	M=10	全フレーム抽出	median	max
Airplane_Flying	0.0114	0.0373	0.420	0.0429	0.0654	0.0742	0.0678	0.0675	0.017	0.141
Boat_Ship	0.0528	0.0324	0.0322	0.0294	0.0231	0.0291	0.0286	0.0276	0.018	0.165
Bus	0.0035	0.0023	0.0035	0.0031	0.0024	0.0029	0.0036	0.0030	0.002	0.032
Cityscape	0.0996	0.1236	0.1250	0.1251	0.1255	0.1216	0.1235	0.1264	0.045	0.21
Classroom	0.0006	0.0017	0.0027	0.0031	0.0048	0.0059	0.0029	0.0032	0.002	0.116

いずれのクラスにおいても，マルチフレームを使用することで従来手法であるキーフレームのみの認識結果を上回った．最も数値が上昇したクラスは Airplane_Flying であり，キーフレームのみの認識に比べて最大 700%上昇した．しかし，フレームの抽出枚数を増やしたにも関わらず適合率が下降したクラスが生じてしまった．また，フレーム選択方法の違いを見ると，M フレーム置きにフレームを選択する手法の方が Boat_Ship を除く 4 クラスで良い結果が得られた．M フレーム置きにフレームを取得する場合，短いショットでは 1 枚しかフレームを抽出できない可能性もあるが，長いショットでは非常に多くの枚数を得られるため一部のショットだけ特徴ベクトルの質が向上する．ただし，本実験の結果では M=15 の値が最も高いクラスが多く，全フレーム抽出の値を上回っている結果も得られた．これらの点から，認識結果の向上には単純な枚数だけでなく，中間のフレームを使わないことでノイズとなる特徴量を適切に回避するといった別の要因も含まれていることが考えられる．

次に，MKL-SVM の実行結果を表 4 に示す．図 6 は表 4 のグラフである．また N 枚抽出について，MKL で学習した特徴量の重みの割合を図 7 に示す． 最も値が上昇したクラスは Classroom であり，MKL-SVM を使ったキーフレームのみの認識に比べて 883%上昇した．Bus を除く 4 つのクラスでは，SVM と比較して MKL-SVM で学習，分類を行った方が最大値が上昇している．しかし，Boat_Ship は MKL-SVM でもフレーム枚数に連れて値が下降している．やはり追加したフレームから有効な特徴量よりもノイズのある特徴量を多く抽出

© 2011 Information Processing Society of Japan

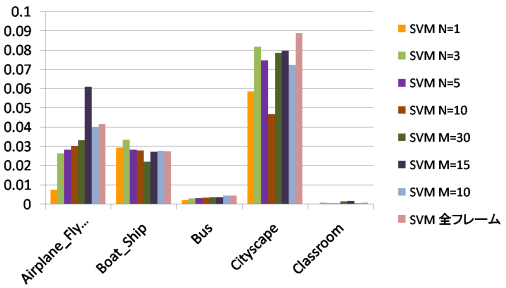

図 5 SVM による結果の比較

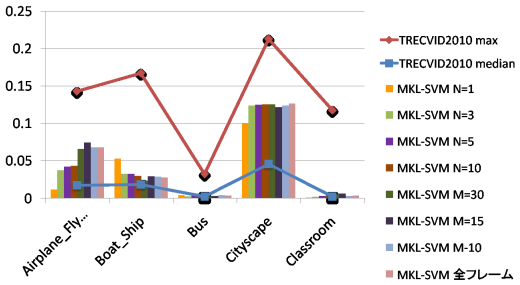

図 6 MKL-SVM による結果の比較

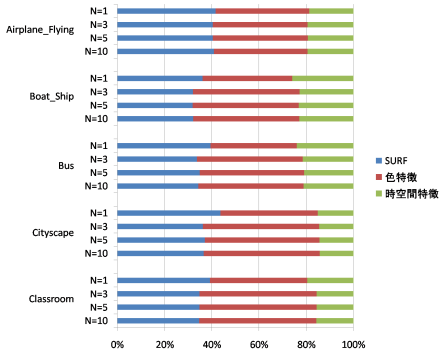

図 7 特徴量の重みの割合 N 枚抽出

してしまったためと考えられる．枚数を変化させたときの重みの変化に注目すると，キーフレームのみである N=1 の処理とマルチフレーム処理では重みの割合が変動していることがわかる．全体的にキーフレームのみの認識よりもマルチフレームの認識の方が色特徴の重みの割合が大きい．これは特徴量の抽出数に関連していると考えられる．本実験では各フレーム画像から SURF を 5000 個，色特徴を全画素から抽出した．TRECVID2010 のショットデータはサイズが 320×240 か，それ以上の横幅を持つ大きさとなっている．よって画素値は 1 枚のフレーム画像から少なくとも 76800 個以上を抽出することができる．Bag-of-Features 表現により量子化は行われるが，複数のフレームを扱ったことで両特徴の抽出数の差が学習した重みに表れたと考えられる．これに対し，マルチフレーム処理で枚数を変化させても割合にほとんど変化はなく，大きく変化しているものでも 1%ほどしか変動はない．

TRECVID2010 の結果と比較すると，最大値には及ばないものの，MKL-SVM での認識結果はいずれのクラスにおいても中央値より良い結果が出ている．SVM での認識結果を見ても，マルチフレームを使った認識は Classroom を除いた 4 クラスで上回った数値を得た．最大値との差が開いている理由としては，実験に使用している特徴量や設定の違いが挙げられる．本実験で使用している特徴量は，動画像認識において基本的な特徴量を最低限しか使用していない．3 章で述べたように，1 位のチームは本研究の倍以上の特徴量を使用しており，さらに codebook のサイズも最大で 4096 次元と設定している．よって，さらなる精度向上のためにはマルチフレーム処理の追求以外に，扱う特徴量の種類を増やす必要がある．

8. ま と め

本研究ではショット認識に対してマルチフレーム手法を適用し，TRECVID2010 の Semantic indexing タスクに則って実験を行うことでその有効性を検証した．ショットから 2 通りの基準でフレームを複数枚選択し，各フレーム画像から SURF および色特徴を，ショット全体から時空間特徴をそれぞれ抽出して学習，分類を行った．評価指標にはランキング付けされた結果の順位を考慮する推定平均適合率を用いた．実験として Semantic indexing タスクの五つの概念を使い，「全体から N 枚を抽出する」，「M フレーム置きに抽出する」と基準を設定したマルチフレーム認識結果をキーフレームのみの認識，全フレームを使ったマルチフレーム認識，TRECVID2010 の全チームの平均結果と比較した．

実験の結果，キーフレームのみの認識に比べて SVM で学習，分類をした場合は推定平均適合率が最大で 700%上昇し，MKL-SVM で学習，分類をした場合は最大で 883%上昇し

© 2011 Information Processing Society of Japan

た．また TRECVID2010 の全チームの平均値と比較した結果，MKL-SVM を使用した設定のいくつかで 5 クラスすべての値を上回る結果を得た．以上の点から，ショット認識におけるマルチフレーム手法の有効性を確認することができた．

9. 今後の課題

実験ではより多くのフレームから特徴量を抽出した場合，逆に精度が悪くなるということが起こった．単純なショット中の位置でフレームを抽出するだけでなく，より有用なフレームを選択できるようにすることも重要である．例えば，時空間特徴の抽出で行ったような選択した他のフレーム (この場合は既に抽出した前のフレーム) との差異を計算した場合に，一定以上特徴量が異なれば新たなフレームとして抽出する，といったことが考えられる．また，そのフレームの差異が大きければ基本的な取得フレーム間隔を小さくし，逆に差異が小さければ間隔を大きくする，つまりフレームの変化によって取得間隔をショット毎に変更するということもできる．

参 考 文 献

1) TRECVID Home Page. *http://www-nlpir.nist.gov/projects/trecvid/*.
2) C.G.M. Snoek, KEA vande Sande, O.deRooij, etal. The mediamill trecvid 2008 semantic video search engine. In *Proc. of TRECVID Workshop*, 2008.
3) D.Cai, X.He, and J.Han. Efficient kernel discriminant analysis via spectral regression. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pp. 427–432. IEEE, 2008.
4) J.Ye, S.Ji, and J.Chen. Multi-class discriminant kernel learning via convex programming. *The Journal of Machine Learning Research*, Vol.9, pp. 719–758, 2008.
5) N.Inoue, S.Hao, T.Saito, K.Shinoda, I.Kim, and C.H. Lee. Titgt at trecvid 2009 workshop. In *Proc. of TRECVID Workshop*, Vol.2, 2009.
6) H.Bay, T.Tuytelaars, and L.VanGool. SURF: Speeded up robust features. In *Proc. of European Conference on Computer Vision*, pp. 404–415, 2006.
7) D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, Vol.60, No.2, pp. 91–110, 2004.
8) 野口顕嗣, 柳井啓司. 動きの連続性を考慮した動画からの局所的な時空間特徴の抽出. In *MIRU*, 2009.
9) B.D. Lucas and T.Kanade. An iterative image registration technique with an application to stereo vision. In *International joint conference on artificial intelligence*, Vol.3, pp. 674–679. Citeseer, 1981.
10) S.Lazebnik, C.Schmid, and J.Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 2169–2178, 2006.
11) S.Sonnenburg, G.Rätsch, C.Schäfer, and B.Schölkopf. Large Scale Multiple Kernel Learning. *The Journal of Machine Learning Research*, Vol.7, pp. 1531–1565, 2006.

付　　録

MKL-SVM，N=10 で実行した上位 15 ショットを示す．赤枠は正解のショットである．

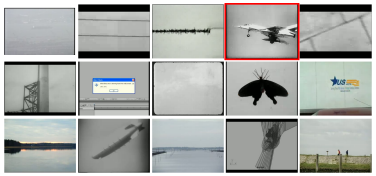

図 8 Airplane_Flying

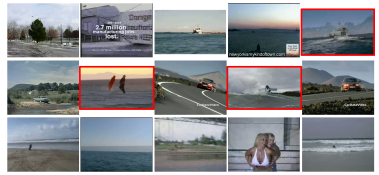

図 9 Boat_Ship

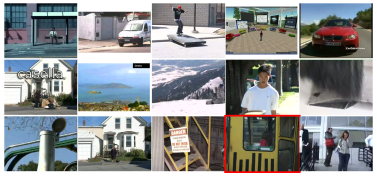

図 10 Bus

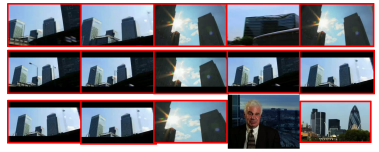

図 11 Cityscape

図 12 Classroom

© 2011 Information Processing Society of Japan