T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	大規模映像資源のためのマルチモーダル高次特徴検出
Title(English)	
著者(和文)	井上 中順, 斉藤 辰彦, 篠田 浩一, 古井 貞熙
Authors(English)	Nakamasa Inoue, Tatsuhiko Saito, Koichi Shinoda, SADAOKI FURUI
出典(和文)	電子情報通信学会論文誌, Vol. J93-D, No. 12, pp. 2633-2644
Citation(English)	, Vol. J93-D, No. 12, pp. 2633-2644
発行日 / Pub. date	2010, 12
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2010 Institute of Electronics, Information and Communication Engineers.

Powered by T2R2 (Science Tokyo Research Repository)

論　文

大規模映像資源のためのマルチモーダル高次特徴検出

井上　中順[†a)]　斉藤　辰彦[††b)]　篠田　浩一[†c)]　古井　貞熙[†d)]

Multimodal High-Level Feature Extraction for Large-Scale Video Resources

Nakamasa INOUE[†a)], Tatsuhiko SAITO[††b)], Koichi SHINODA[†c)], and Sadaoki FURUI[†d)]

あらまし　本研究では，映像の中から「飛行機」や「歌っている人」といった高次特徴を検出するタスクに対し，SIFT 特徴と MFCC 特徴の混合ガウス分布 (GMM) を用いた統計的手法を提案する．検出手法には，話者認識などで用いられてきたゆう度比による検出と，GMM Supervector SVM (GS-SVM) による検出の二つを用いる．ゆう度比による検出では，高次特徴が出現する部分としない部分の GMM をそれぞれ学習し，二つのモデルから得られるゆう度の比をもとに高次特徴を検出する．GS-SVM では，各ショットに対する GMM を求め，GMM 間の距離から定義される RBF カーネルを用いた SVM で学習・識別を行う．最後に，各手法から対数ゆう度比を求め，その重み付き和により手法の融合を行う．TRECVID2009 のデータセットを用いて評価実験を行った結果，Mean Average Precision は SIFT 特徴と GS-SVM を用いた場合の 0.141 から，融合手法により 0.173 まで向上した．

キーワード　高次特徴検出，SIFT，MFCC，ゆう度比，GMM Supervector SVM

1. まえがき

近年，ディジタル録画機器の発展やネットワークの高速化に伴い，個々のユーザが膨大な映像データにアクセスできるようになった．更に，映像データは日々増加し続けており，膨大なデータベースの中から必要な情報を効率的に検索する技術が切望されている．

映像検索においては，映像中の何らかの「意味」をもつものが検索の対象となる．例えば，「飛行機」，「船」などの物体，「歌っている」，「握手している」などのイベント，「夜景」，「街並み」などの風景（シーン）などである．ここでは，これらをまとめて，画像から直接得られる画像特徴（低次特徴）と区別して，「高次特徴」と呼ぶ．本論文では映像からこれらの高次特徴を検出することを目的とする．

一般に，映像の信号処理により得られる低次特徴と検索対象としての高次特徴との間には明示的な関係はなく，その関係性を獲得することは容易ではない（セマンティックギャップ）．そこで，現在の高次特徴検出の研究では，高次特徴のラベルが付与された学習サンプルを利用して，映像と高次特徴の関係を表す統計モデルを構築する研究が行われている．

従来，この統計的高次特徴検出の研究では，画像処理における一般物体認識の技術がしばしば用いられてきた．そこでは，まず，画像から，Scale-Invariant Feature Transform (SIFT) [2] や Speeded Up Robust Features (SURF) [3] などの局所特徴を抽出し，量子化を行う．次に，文書検索で用いられる Bag of Words を応用した，Bag of Visual Words (BoW) [4] を構築し，Visual Word の分布に対する統計モデルを用いて，検出を行う．更に，Visual Word 間の共起を用いる方法，前景と後景の違いを利用する方法，画像中の位置情報を利用する方法など，様々な拡張が試みられている．これらの手法の多くは，限られた条件下では高い性能をもつことが報告されているが，一般のテレビ番組などを対象とした映像検出に適応した際，十分な性能が得られるとは限らない．BoW のような，

† 東京工業大学大学院情報理工学研究科計算工学専攻，東京都
Department of Computer Sience, Tokyo Institute of Technology, 2–12–1 Ookayama, Meguro-ku, Tokyo, 152–8552 Japan
†† 東京工業大学大学院総合理工学研究科物理情報システム専攻，横浜市
Department of Information Processing, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama-shi, 226–8503 Japan
a) E-mail: inoue@ks.cs.titech.ac.jp
b) E-mail: saito.t.aa@m.titech.ac.jp
c) E-mail: shinoda@cs.titech.ac.jp
d) E-mail: furui@cs.titech.ac.jp

一般物体認識の技術を映像検出に用いる際の課題として大きなものに以下の二つがある．

まず，映像は画像の集合であり，得られる画像のサンプル数は大幅に多い．現在はキーフレームと呼ばれる一部の画像のみを用いる手法が主流であるが，より多くのサンプルを用いた方が，性能が高くなることが期待できる．反面，それらの多くの画像すべてに対し，正確なラベル付けを行うことは困難である．一方，協調的タグ付け (collaborative annotation) [5] を用いれば，タグ付けを行う人により揺らぎがあるものの，安価かつ大量のラベルを得られる．よって，このようなラベルを用いても，頑健に学習ができるモデルが必要となる．

次に，映像に付随する音響・音声情報の利用が挙げられる．過去のいくつかの研究（例えば [6], [7]）で，音響・音声情報が高次特徴の検出に有効であることが報告されているが，音響・音声情報に限った評価であったり，あるいは，比較的小規模な評価データを用いた評価であることが多く，今回対象とする大規模な映像データにおける，映像と組み合わせた場合の性能は必ずしも明らかではない．

本論文は，ビデオショットに対し高次特徴のラベルが付いた大量の映像データを対象とし，動画像特徴と音響特徴の両方を入力とした，頑健なマルチモーダル高次特徴検出手法を提案する [1]．このデータにおいては，ショット中に高次特徴のない画像も存在し，また，高次特徴が存在している場合も，画面中央や視覚注意を引く領域にあるとは限らないため，物体の位置情報などを利用した手法の適用が難しい．

そこで，本手法では，映像から得られる動画像・音響特徴をそれぞれ混合ガウス分布 (GMM) でモデル化し，ゆう度比と GMM Supervector SVM (GS-SVM) [8] による検出を行う．本手法では，GMM によりモデル化を行うため，従来の BoW に比べ，対象の見え，欠落などにより生じる揺らぎに対して頑健な手法となっている．また，動画像特徴はショット中の全フレーム画像から抽出し，データ量が多いときに GMM によるモデル化が有効であることを示す．最後に，各手法から対数ゆう度比を求め，その重み付き和により融合を行うことで，動画像・音情報を総合的に加味したシステムを実現する．

本論文の構成は以下のとおりである．**2.** で関連研究について述べる．**3.** で特徴抽出，**4.** で検出手法をそれぞれ説明する．**5.** では TRECVID2009 [9], [10] の映像コーパスを用いた評価実験の結果を示し，**6.** で結論と今後の課題について述べる．

2. 関連研究

高次特徴検出に対するアプローチとして代表的であるのは，画像から抽出した局所特徴を量子化した Visual Word のヒストグラムによって画像を識別する BoW [4] である．BoW の拡張として，局所領域抽出・局所特徴記述・クラスタリングの 3 点に関して様々な手法が提案されている．

局所領域抽出としては，LoG [11] オペレータの極大点から抽出を行う方法やその近似の DoG [2] オペレータを用いる方法がある．また，コーナ点に注目した Harris-Laplace やヘッセ行列を用いた Hessian-Laplace も提案されている [12]．これらの抽出手法はスケール変化に対して不変であるため，検出対象の高次特徴の大きさが変化しても頑健な認識が可能である．更に，アフィン変換に頑健な Harris-Affine と Hessian-Affine も提案されている [13]．Peng ら [14] は 6 種類の抽出手法を組み合わせて用いることで，認識精度が向上することを示している．

局所特徴としては，SIFT [2] と SURF [3] が最も一般的である．SIFT と SURF では，輝度情報のみを用いて特徴量記述を行うため，色相と彩度に特徴のある高次特徴の抽出には適してない．そこで，最近では，色情報を取り込んだ Color-SIFT [15] なども提案されている．

クラスタリングに関しては，k-means 法が最もよく用いられており，ソフトクラスタリングによる BoW ヒストグラムの作成（例えば Kernel Codebooks [16]）が注目されている．ソフトクラスタリングでは，ハードクラスタリングに見られるスパース性が失われるため高速化には向かないが，量子化誤差に関する問題を解決できるため，認識精度の向上が見られる．また，GMM を用いてソフトクラスタリングを行う BoW [17] や，GMM で求めた事後確率を特徴ベクトルとして用いる手法 [18] も提案されている．

映像からの高次特徴検出に対して，BoW のような画像認識手法を用いる場合は，ショットの中からそれを代表する 1 枚のフレーム画像（キーフレーム）を抽出するということが行われてきた．しかし，最近では複数のフレーム画像を用いるマルチフレームの効果が大きいことが確認されている．例えば，Snoek ら [19] はショットから 10 枚のフレーム画像を抽出して BoW

のヒストグラムを作成することで認識精度が向上することを示している．

映像中の音情報を利用した研究としては，動画像特徴と音響特徴を結合した特徴量を用いる手法 [6] が提案されている．また，音声認識結果を利用した研究もある [7]．

一方，音声を入力とした話者認識の手法として，ゆう度比による検出や GMM Supervector SVM [8]（以下 GS-SVM）による検出が提案されており，これらの手法は映像からの高次特徴の検出にも応用できると考えられる．特に GS-SVM と SIFT 特徴の組合せのイベント認識に対する有用性は Zhou ら [20] により報告されているが，音響的特徴との融合は行われていない．そこで，本研究では，ゆう度比検出，GS-SVM に SIFT 特徴と MFCC 特徴を組み合わせ，マルチモーダルな高次特徴検出を実現する．更に，大規模な映像データに対する本手法の有用性を示す．

3. 特徴抽出

3.1 動画像特徴

動画像特徴には SIFT 特徴 [2] を用いる．SIFT は局所領域の抽出と特徴量の記述を行うアルゴリズムである．明るさの変化やアフィン変換に頑健な特徴量が得られるため，照明や視点の変化に頑健な高次特徴検出が期待される．SIFT 特徴は 1 枚の画像から局所領域の数（一般に数百個）だけ抽出される．特徴ベクトルは，16 ブロックに分割された局所領域における 8 方向の輝度こう配ヒストグラムを表しており，$16 \times 8 = 128$ 次元である．本研究では，SIFT 特徴を主成分分析により 32 次元に圧縮したものを用いる．局所領域抽出に関しては，アフィン変換に頑健な Harris-Affine [13] と Hessian-Affine [13] の 2 種類を別々に用いる．

3.2 音響特徴

音響特徴には MFCC 特徴を用いる．MFCC 特徴は元来，音声認識向けに開発された特徴量であるが，一般の音響分類にも有用である．

本研究では 12 次元の MFCC に，1 階・2 階微分である ΔMFCC，ΔΔMFCC，対数パワーの 1 階・2 階微分である Δ 対数パワー，ΔΔ 対数パワーを合わせた 38 次元の特徴量を用いる．

4. 検出手法

この章では，ゆう度比による検出と GS-SVM による検出についてそれぞれ説明し，最後に融合手法について述べる．融合手法では，SIFT/MFCC 特徴とゆう度比/GS-SVM の組合せによる 4 通りの手法を融合する．以下では，第 s 番目のショットから得られた SIFT/MFCC 特徴を X_s，検出対象の高次特徴が出現しているときに $+1$，そうでないとき -1 をとる確率変数を H_s とする．

4.1 ゆう度比による検出

ゆう度比による検出では，二つのゆう度の比である

$$L_s = \frac{p(X_s|H_s=+1)}{p(X_s|H_s=-1)} \tag{1}$$

によって高次特徴の検出を行う．

確率分布は $X_s = \{x_i\}_{i=1}^{n_s}$ が i.i.d. であると仮定し，GMM を用いて推定する．GMM の確率密度関数は以下で与えられる．

$$p(x|\theta) = \sum_{k=1}^{K} w_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{2}$$

ここで，K は混合数，$\theta = \{w_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ は GMM のパラメータであり，w_k は混合重み，$\mathcal{N}(x|\mu_k, \Sigma_k)$ は平均ベクトル μ_k，分散行列 Σ_k のガウス分布の確率密度関数を表す．

GMM のパラメータ θ は EM アルゴリズムを用いた最ゆう法で推定する．高次特徴が出現するショット全体から $\hat{\theta}^{(\mathrm{HLF})}$ を，出現しないショット全体から $\hat{\theta}^{(\mathrm{UBM})}$ をそれぞれ推定し，以下のようにゆう度比を計算する．

$$L_s = \frac{p(X_s|\hat{\theta}^{(\mathrm{HLF})})}{p(Xs|\hat{\theta}^{(\mathrm{UBM})})} = \prod_{i=1}^{n_s} \frac{p(x_i|\hat{\theta}^{(\mathrm{HLF})})}{p(x_i|\hat{\theta}^{(\mathrm{UBM})})} \tag{3}$$

ここで，本来 $\hat{\theta}^{(\mathrm{UBM})}$ は高次特徴ごとに別々のものを用意する必要があるが，全体のショットに対して高次特徴が出現する割合が少ない場合，どの高次特徴も出現しない部分から $\hat{\theta}^{(\mathrm{UBM})}$ を推定することで，共通のパラメータを用いることができる．また，$p(X_s|H_s=-1)$ の代わりに $p(X_s)$ を用いてもよい．その場合，映像全体から特徴をサンプリングするため，得られたモデルは Universal Background Model (UBM) と呼ばれる．TRECVID のデータセットは上記の仮定を満たしており，どちらの方法で $\hat{\theta}^{(\mathrm{UBM})}$ を推定してもほぼ同様の結果が得られる．

4.2 GS-SVM による検出

GMM Supervector SVM (GS-SVM) [8] とは，GMM のパラメータから得られる supervector を利用した SVM で学習・識別を行う手法である．

ゆう度比による検出では，高次特徴ごとにGMMを求めるのに対し，ここではショットごとにGMMを求める．しかし，時間が短いショットではパラメータの推定に十分な量のSIFT/MFCC特徴が得られない可能性があるため，最大事後確率 (Maximum A Posteriori; MAP) 適応によってパラメータを推定する．

MAP適応では，パラメータ θ の事前分布 $p(\theta)$ を導入し，以下のような基準で第 s 番目のショットに対するパラメータ $\hat{\theta}^{(s)}$ を求める．

$$\hat{\theta}^{(s)} = \operatorname*{argmax}_{\theta} \left(\sum_{i=1}^{n_s} \log p(x_i|\theta) + \log p(\theta) \right) \quad (4)$$

ここで，事前分布はUBMのパラメータ $\hat{\theta}^{(\mathrm{UBM})}$ をもとに決定する．本研究では，ショットごとに平均ベクトルを1回更新し，ショットに対するGMMの推定を行う．具体的には，$\{\mu_k\}_{k=1}^{K}$ が互いに独立だと仮定し，

$$p(\mu_k) = \mathcal{N}(\mu_k|\mu_k^{(\mathrm{UBM})}, \tau\Sigma_k^{(\mathrm{UBM})}) \quad (5)$$

を事前分布として用いることで，

$$\hat{\mu}_k^{(s)} = \frac{\tau\mu_k^{(\mathrm{UBM})} + \sum_{i=1}^{n_s} c_{ik}x_i}{\tau + C_k} \quad (6)$$

という更新式を得る．ここで，

$$c_{ik} = \frac{w_k\mathcal{N}(x_i|\mu_k^{(\mathrm{UBM})}, \Sigma_k^{(\mathrm{UBM})})}{\sum_{k=1}^{K} w_k\mathcal{N}(x_i|\mu_k^{(\mathrm{UBM})}, \Sigma_k^{(\mathrm{UBM})})}, \quad (7)$$

$$C_k = \sum_{i=1}^{n_s} c_{ik} \quad (8)$$

であり，τ は事前分布への依存度を決めるハイパパラメータである．

次に，第 s，t 番目のショットに対応するGMMのパラメータ $\hat{\theta}^{(s)}$，$\hat{\theta}^{(t)}$ からショット間の距離 $d(s,t)$ を定める．パラメータの更新は式 (6) で行われているため，混合成分の対応が取れていることに注意すれば，対応する混合成分ごとの距離の重み付き和からGMM間の距離を定めることができる．具体的には，混合成分間のマハラノビス距離の平方の重み付き和から以下のように距離 $d(s,t)$ を定める．

$$d(s,t) = \sum_{k=1}^{K} w_k^{(\mathrm{UBM})}(\mu_k^{(s)} - \mu_k^{(t)})^{\mathrm{T}}(\Sigma_k^{(\mathrm{UBM})})^{-1}(\mu_k^{(s)} - \mu_k^{(t)}) \quad (9)$$

更に，$||\phi(s) - \phi(t)||_2^2 = d(s,t)$ を満たすような特徴写像 ϕ も以下のように構成することが可能である．

$$\phi(s) = (v_1^{\mathrm{T}}, v_2^{\mathrm{T}}, \cdots, v_K^{\mathrm{T}})^{\mathrm{T}} \quad (10)$$

$$v_k = (w_k^{(\mathrm{UBM})}\Sigma_k^{(\mathrm{UBM})})^{-\frac{1}{2}}(\mu_k^{(s)} - \mu_k^{(\mathrm{UBM})}) \quad (11)$$

この $\phi(s)$ が第 s 番目のショットに対応するGMMのsupervectorである．また，その次元は特徴量の次元 d と混合数 K の積で与えられる．

カーネル関数 $k(s,t)$ には，以下で定義されるRBFカーネルを用いる．

$$k(s,t) = \exp\left(-\gamma||\phi(s) - \phi(t)||_2^2\right) \quad (12)$$

ここで，γ は実験により決定される定数である．

最後に，事後確率 $p(H_s = +1|X_s)$ を求めるため，SVMの出力を用いた事後確率推定を行う．事後確率は，SVMの識別関数 f からシグモイドフィッティングにより求める [21]．

4.3 融合手法

ゆう度比とGS-SVMによる検出の結果の融合は対数ゆう度比 ℓ_s の重み付き和によって行う．

まず，ゆう度比による検出では，対数ゆう度比 ℓ_s を直接

$$\ell_s = \log L_s \quad (13)$$

によって求める．次に，GS-SVMに関しては

$$\begin{aligned} \ell_s &= \log \frac{p(X_s|H=+1)}{p(X_s|H=-1)} \\ &= \log \frac{p(H=+1|X_s)}{p(H=-1|X_s)} \cdot \frac{p(H=-1)}{p(H=+1)} \\ &= \log \frac{p(H=+1|X_s)}{1-p(H=+1|X_s)} + \text{const.} \end{aligned} \quad (14)$$

と計算できるので，事後確率のオッズ比の対数（第1項）を対数ゆう度比の代わりに用いる．

最後に，SIFT/MFCC特徴とゆう度比/GS-SVMの計4通りの組合せから求めた対数ゆう度比をそれぞれ $\ell_s^{(\text{SIFT-LR})}$，$\ell_s^{(\text{SIFT-SVM})}$，$\ell_s^{(\text{MFCC-LR})}$，$\ell_s^{(\text{MFCC-SVM})}$ として，融合対数ゆう度比 ℓ_s を次式で求める．

$$\ell_s = \sum_{\substack{\mathrm{F}\in\{\text{SIFT},\text{MFCC}\} \\ \mathrm{M}\in\{\text{LR},\text{SVM}\}}} w^{(\mathrm{F\text{-}M})}\ell_s^{(\mathrm{F\text{-}M})} \quad (15)$$

ここで，手法に対する重み $w^{(\mathrm{F\text{-}M})}$ はクロスバリデーション (CV) で高次特徴ごとに最適なものを決定する．

5. 評価実験

5.1 実験条件

評価実験では，TRECVID2009 の development data として提供されている 100 時間の映像を 2 分割し，学習・テスト用に固定した．映像はオランダで放送された教育・ドキュメンタリー番組が主となっている．データセットには，ショットの境界と各ショットに出現する高次特徴の正解ラベルが与えられている．ショットの総数は，学習用が 18,120，テスト用が 18,142 である．検出単位はショット（カメラの切換わりがないフレームの集合）であり，検出結果は順位付けを行って出力する．

検出対象となる高次特徴は TRECVID2009 で用いられた 20 種類である（図 1，詳細は付録を参照）．これらは，映像検索で用いられることを想定して選ばれており，出現頻度は中程度（データセット中で 100 から 500 回程度の出現）を目安としている．高次特徴の出現数は図 1 のとおりで，検出対象の高次特徴が出現するショットは平均で全体の 0.92%である．20 種類の中には，Boat_Ship などの物体をはじめ，People-dancing，Person-eating といった動作も含まれている．音響に特徴があるものとしては，Singing，Person-playing-a-musical-instrument が挙げられる．ラベル付けはキーフレーム画像を用いた協調的アノテーションシステムで行われている．そのため，例えば Singing では，マイクをもたずに歌っているショットなどが見落とされている可能性もある．また，Traffic-intersection，Person-riding-a-bicycle，Person-eating，Hand はテレビ番組のオープニングとエンディングに出現があり，学習データ・テストデータにほぼ同一のショットが含まれている．

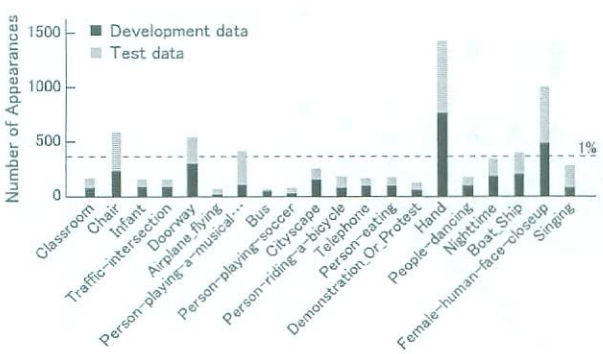

図 1　高次特徴の出現数
Fig. 1　Number of appearances of high-level features.

SIFT 特徴の抽出には Mikolajczyk らのツール [13] を利用し，局所領域には Harris-Affine と Hessian-Affine を別々に用いた．局所領域の数は表 1 のとおりである．GMM のパラメータ推定には音声認識のツールキットである HTK [22] を利用した．UBM の推定では，100 万個の特徴をランダムにサンプリングして用いた．また，分散行列は対角成分のみを推定し，MAP 適応におけるパラメータ τ は HTK の標準である 20.0 とした．SVM による事後確率推定には libSVM [23] を用いた．RBF カーネルのパラメータ γ は平均距離の逆数とし，学習の際にはすべての負例を用いた．重み係数 $w^{(\mathrm{F\text{-}M})}$ は学習データに対する 2-fold cross validation (CV) により決定した．その際，重みは 0.01 刻みで最適化を行った（四つ以上の手法を融合する際には，0.05 刻みで最適化を行って得られた重みの周辺を 0.01 刻みで再探索した）．

5.2 評価方法

検出結果の評価は，TRECVID における評価基準と同様，各高次特徴に対する Average Precision (AP) の単純平均である Mean AP により行う．AP は順位付きの検索結果を評価する際に用いられる評価尺度で，Precision-Recall 曲線と座標軸に囲まれた部分の面積に相当する．また，計算式は以下のとおりである．

$$\mathrm{AP} = \frac{1}{R}\sum_{r=1}^{N}\mathrm{Pr}(r)\mathrm{Rel}(r) \tag{16}$$

ここで，R は正解の総数，N は検索結果の総数，$\mathrm{Pr}(r)$ は第 r 位までの検出結果における Precision，$\mathrm{Rel}(r)$ は第 r 位の検出結果が正解であった場合に 1，そうでないときに 0 をとる関数である．また，TRECVID では検出結果の上位 2000 ショットまでを提出するため，ここでは $N = 2000$ として評価を行った．

5.3 実験結果

5.3.1 各手法に対する結果

図 2 に各手法に対する Mean AP を示す．SIFT-LR，SIFT-SVM，MFCC-LR，MFCC-SVM はそれぞれ，SIFT/MFCC 特徴とゆう度比 (LR)/GS-SVM (SVM) を用いた際の結果である．SIFT 特徴は Harris-Affine 領域から抽出した．横軸は混合数 K であり，$K =$

表 1　フレーム/ショット当りの平均局所領域数
Table 1　Number of local regions per frame/shot.

局所領域	領域数(個/フレーム)	領域数(個/ショット)
Harris-Affine	304.7	80557.8
Hessian-Affine	276.9	73200.5

$1, 2, 4, \cdots, 512$ と変化させた．また，$K = 512$ のときの各高次特徴に対する AP を図 3 に示す．

四つの手法の中で SIFT-SVM が最も高い性能を示している．混合数は 512 が最も良い．また，Mean AP はまだ上昇傾向にあり，更に混合数を大きくすることで精度の改善が期待される．SIFT-LR は次に高い性能を示しているが，どの高次特徴に対しても SIFT-SVM の方が良く，Mean AP にも大きな差がある．SIFT-LR では，UBM と各高次特徴の GMM で共通の混合数を用いているが，情報量基準などを用いて UBM 及び高次特徴ごとに最適な混合数を決定することも可能である．

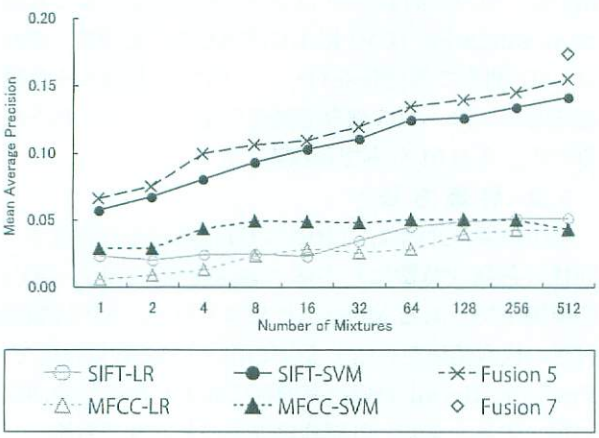

図 2　各手法に対する Mean AP
Fig. 2　Comparison of Mean APs for different schemes.

MFCC 特徴を用いた手法は，特に，Singing に関して効果が非常に大きい．MFCC-SVM と MFCC-LR を比較した場合は，MFCC-SVM の方が全体的に高い精度を示しているが，混合数が大きい部分では同程度の結果が得られている．混合数に関しては，256，512 の二つで大きな差がなく，これ以上混合数を大きくする必要はないといえる．また，高次特徴ごとに最適な混合数が異なるため，それを決定する手法も必要だと考えられる．高次特徴の出現数と最適な混合数の相関については，**5.3.7** で考察する．

5.3.2　BoW との比較

比較実験として，BoW [4] を用いて高次特徴検出を行った結果を表 2 に示す（表中の Fusion に関しては後述）．第 2 列には有意水準 5%の Randomization test [24] の結果を示した．数字は各手法が他の手法よりも有意となった数である．例えば，BoW の「2」は，下から二つの MFCC-SVM，MFCC-LR と有意差があり，残る SIFT-LR とは有意差がないことを表している．

BoW では，特徴量に Harris-Affine 領域から抽出した SIFT 特徴を用いた．辞書サイズは混合数に合わせて 512 とした（BoW の辞書サイズも 1 から 512 の中

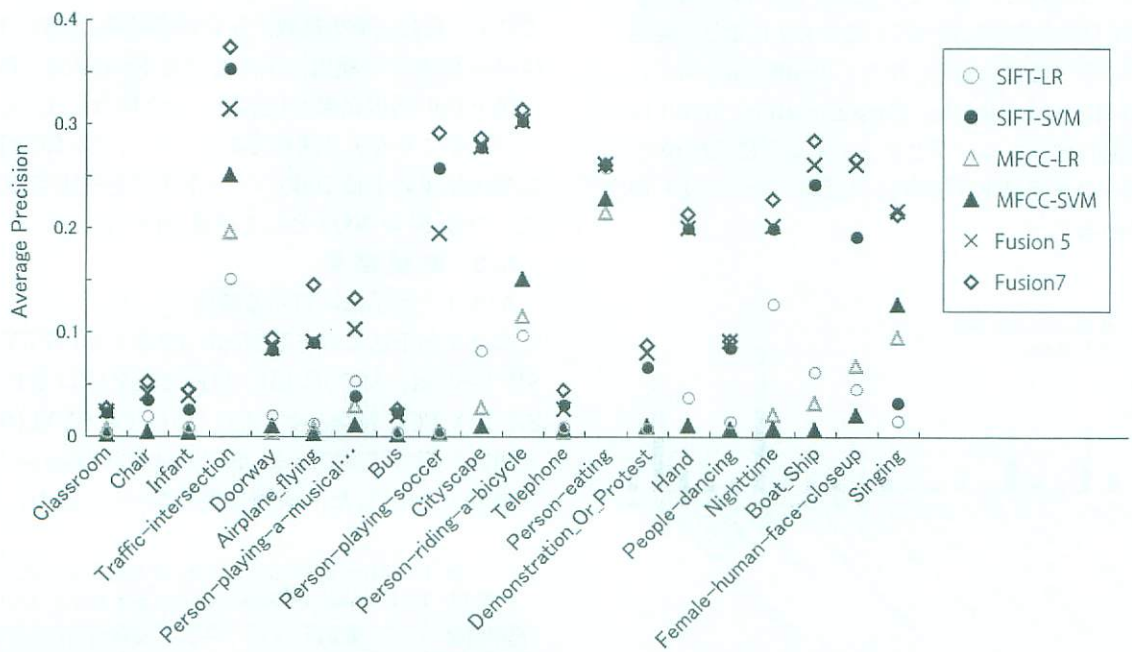

図 3　各高次特徴に対する AP（混合数 $K = 512$）
Fig. 3　APs for each high-level feature ($K = 512$).

表 2 各手法に対する Mean AP と Singing, Female-human-face-closeup (Female), Person-playing-a-musical-instrument (M.I.) の AP (混合数 $K = 512$). 第 2 列は有意水準 5%の Randomization test の結果で, 数字は各手法が他の手法よりも有意となった数である.

Table 2 Mean APs for different schemes ($K = 512$). R denotes a result of Randomization test ($p = 0.05$).

手法	R	Mean AP	Singing	Female	M.I.
Fusion 8	14	0.186	0.233	0.271	0.149
Fusion 7	13	0.173	0.213	0.266	0.132
Fusion 6	10	0.155	0.038	0.206	0.048
Fusion 5	9	0.154	0.216	0.261	0.102
Fusion 4	8	0.145	0.151	0.205	0.022
SIFT-SVM	8	0.141	0.032	0.192	0.038
Fusion 1	7	0.138	0.037	0.193	0.063
$\mathrm{SIFT_{hes}}$-SVM	7	0.129	0.025	0.163	0.044
BoW (Multiframe)	6	0.097	0.015	0.115	0.041
Fusion 3	3	0.064	0.113	0.100	0.096
BoW	2	0.060	0.004	0.049	0.020
Fusion 2	2	0.050	0.163	0.066	0.016
SIFT-LR	0	0.051	0.015	0.045	0.053
MFCC-SVM	0	0.043	0.126	0.020	0.010
MFCC-LR	0	0.042	0.095	0.067	0.028

では 512 の場合が最も性能が良い). SVM のカーネルは Zhang ら [25] により高い識別能力が報告されている χ^2RBF カーネルとした. また, パラメータ γ には平均距離の逆数を用い, 学習の際にはすべての負例を用いた. 表 2 の BoW はショットから 1 枚のフレーム画像を抽出した際の結果, BoW (Multiframe) は複数のフレーム画像（ここでは最大 63 フレーム）から抽出した際の結果である. BoW (Multiframe) では, 複数のフレーム画像から抽出した SIFT 特徴の和集合からヒストグラムを作成した. マルチフレームの効果に関しては **5.3.5** で詳しく述べる.

5.3.3 融合手法

融合手法として, 以下の 5 通りについて実験を行った.

Fusion 1 : SIFT-LR + SIFT-SVM
Fusion 2 : MFCC-LR + MFCC-SVM
Fusion 3 : SIFT-LR + MFCC-LR
Fusion 4 : SIFT-SVM + MFCC-SVM
Fusion 5 : SIFT-LR + SIFT-SVM
+ MFCC-LR + MFCC-SVM

各融合手法に対する Mean AP を表 2 に示す.

更に, 音響との融合の効果が大きかったものとして, Singing, Female-human-face-closeup, Person-playing-a-musical-instrument の三つの高次特徴に対する AP も同表に示した. また, Fusion 5 に関しては, 図 2, 図 3 にも結果をプロットした.

まず, Fusion 1, 2 はゆう度比と GS-SVM の組合せである. Fusion 2 の結果は融合前よりも有意によく, 融合の効果が確認された. しかし, Fusion 1 と SIFT-SVM には有意差がなく, Mean AP の値もわずかではあるが低下している. これは, CV による重み決定がうまくいかなかったことが原因で, SIFT 特徴については GS-SVM のみを用いてもよいと考えられる.

次に, Fusion 3, 4, 5 は SIFT 特徴と MFCC 特徴を組み合わせたものである. Fusion 3 では融合前の手法と比べ有意差が得られたが, Fusion 4, 5 については得られなかった. しかし, 表に挙げた三つの高次特徴では音響との融合の効果が大きく現れている. Female-human-face-closeup では, 女性がしゃべっている場面が多く, 音響特徴により男女の区別ができた.

5.3.4 Harris-Affine 領域と Hessian-Affine 領域

次に, SIFT 特徴抽出における局所領域を Harris-Affine 領域から Hessian-Affine 領域に変えて実験を行った. 先の結果を踏まえ, 検出には GS-SVM を用い, 混合数は 512 とした. 結果を表 2 の $\mathrm{SIFT_{hes}}$-SVM に示す. また, 以下の融合手法で実験を行った.

Fusion 6 : SIFT-SVM + $\mathrm{SIFT_{hes}}$-SVM
Fusion 7 : Fusion 5 の 4 手法 + $\mathrm{SIFT_{hes}}$-SVM

更に, 参考用として, Fusion 7 において重み係数をテストデータのラベルを用いて事後的に決定した Fusion 8 も用意した.

Harris-Affine 領域と Hessian-Affine 領域を比べると, 前者の方が良いという結果であり, これらを合わせた Fusion 6 では融合前の SIFT-SVM, $\mathrm{SIFT_{hes}}$-SVM との有意差が得られた. Harris-Affine 領域ではコーナ点を中心に局所領域が抽出されるのに対し, Hessian-Affine 領域ではフラットな部分も局所領域として抽出されるため, 2 種の領域が相補的な働きをして精度の向上につながったと考えられる. 更に, 音響特徴も合わせた Fusion 7 では Mean AP が 0.173 まで向上した.

一方, Fusion 8 の結果より, 重みを適切に選べば Mean AP が最大 0.186 となることが分かる. 現在は 2-fold CV により重みを決定しているが, 重みの決定には工夫の余地があるといえる.

5.3.5 マルチフレームの効果

SIFT-SVM と BoW について, 特徴抽出を行うフレーム数を変えて実験を行った結果を図 4 に示す. ここで, フレームは, ショットを 2^n 等分する際に境界と

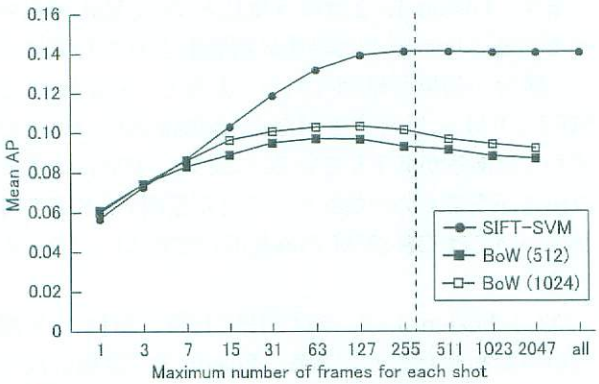

図 4 フレーム数に対する Mean AP の変化．破線はショット中のフレーム数の平均値を表す．
Fig. 4 Mean APs with different numbers of frames. The dotted line indicates the average number of frames in a shot.

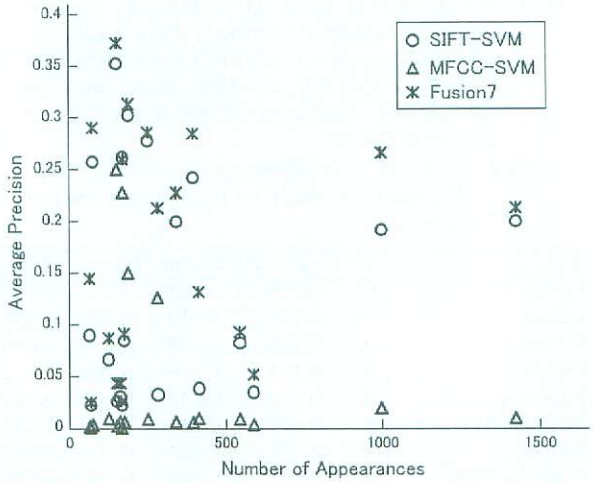

図 5 高次特徴の出現数と Average Precision の関係
Fig. 5 Relationship between the number of appearances and AP.

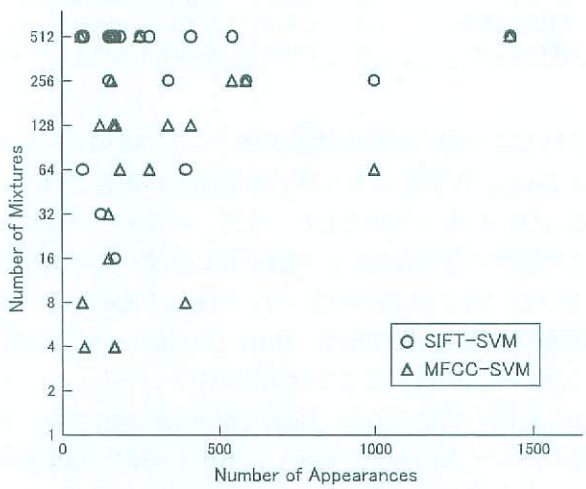

図 6 高次特徴の出現数と最適な混合数の関係
Fig. 6 Relationship between the number of appearances and the optimal number of mixtures.

なる部分を選んだ．例えば，$n = 1$ のときはショット中央のフレームのみを用いている．

図 4 より，特徴抽出を行うフレーム数が多い部分で，SIFT-SVM が BoW よりも良い性能を示していることが分かる．よって，SIFT-SVM はデータ量が多い場合に適した手法であるといえる．この結果は，二つの手法における SIFT 特徴の分布の推定法の違い（BoW ではヒストグラム，SIFT-SVM では GMM を用いている）によるものだと考えられる．また，SIFT-SVM では，全フレーム画像を用いた場合と 127 枚の場合で同程度の結果が得られている．ショットは平均 270 枚のフレーム画像を含むため，ここでは約半数を用いればよいということである．

更に，BoW に関しては辞書サイズを 1024 とした際の結果も同図に示した．辞書サイズ 512 と 1024 を比較すると，1024 の方がどのフレーム数に関しても良い結果を示している．これより，BoW に関しても，SIFT-SVM と同様，辞書サイズを更に大きくすることで，認識精度の向上が期待される．

5.3.6 高次特徴の出現数と Average Precision

高次特徴の出現数と AP の相関を図 5 に示す．図には SIFT-SVM，MFCC-SVM，Fusion 7 の三つの場合が示されており，相関係数はそれぞれ 0.094，−0.182，0.134 であった．

高次特徴の出現数と AP にはほとんど相関が見られない．これより，高次特徴の種類により検出の難しさが大きく異なると考えられる．例えば，Airplane_flying はデータ量が非常に少ないにもかかわらず高い精度が得られている．Airplane_flying は，「飛行機の背景に空がある場面」という限られたショットのみが検出の対象となり，カテゴリー内における多様性が少ない．そのため，検出が容易になったと考えられる．一方，Chair はカテゴリー内の多様性が非常に多いことと，椅子以外のものが画面に多く映っていることが検出を困難にしている．実際，データ量が 20 種の高次特徴の中で 4 番目に多いにもかかわらず，AP は 0.05 と低い．現在，高次特徴検出の総合評価は Mean AP で行っているが，今後は高次特徴ごとの検出の難しさを考慮した評価尺度も必要となるであろう．

5.3.7 高次特徴の出現数と最適な混合数

高次特徴の出現数に対する最適な混合数の値を図 6

表 3　Fusion 4 における音・動画像の重み
Table 3　Weight coefficients between audio and visual in the Fusion 4.

High-level Feature	MFCC-SVM	SIFT-SVM
Singing	0.78	0.22
Person-playing-a-musical-...	0.72	0.28
Person-eating	0.70	0.30
Traffic-intersection	0.67	0.33
Hand	0.63	0.37
Telephone	0.51	0.49
Person-playing-soccer	0.46	0.54
Demonstration_Or_Protest	0.33	0.67
Bus	0.25	0.75
Female-human-face-closeup	0.24	0.76
Infant	0.24	0.76
Classroom	0.18	0.82
Airplane_flying	0.17	0.83
Boat_Ship	0.12	0.88
Chair	0.06	0.94
Doorway	0.05	0.95
Nighttime	0.03	0.97
Person-riding-a-bicycle	0.02	0.98
Cityscape	0.01	0.99
People-dancing	0	1.00

に示す．ここで，最適な混合数とは，AP が最大になった混合数である．相関係数は SIFT-SVM，MFCC-SVM でそれぞれ 0.140，−0.014 と，ここでも強い相関はない．

また，同図より SIFT-SVM の方が MFCC-SVM よりも大きい混合数を必要としていることが分かる．これは，各ショットから得られる SIFT 特徴の数が MFCC 特徴に比べて多いためである．更に，SIFT-SVM と MFCC-SVM における最適な混合数間の相関係数を計算したところ，0.525 となり中程度の相関があることが分かった．例えば，Nighttime では暗くて静かな場面が多いため，SIFT・MFCC 両者とも小さい混合数でモデル化することができると考えられる．

5.3.8　音響と動画像の重み

各高次特徴に対する音響と動画像の重みを表 3 に示す．この結果は，音響に MFCC-SVM，動画像に SIFT-SVM を用いた Fusion 4 のものであり，音響の重みが大きい順に並べている．

Singing と Person-playing-a-musical-instrument に関しては音響の重みが大きい．これらの高次特徴は音楽に関係しており，音の情報の重要性を確認することができる．一方，Person-eating，Traffic-intersection，Hand は，テレビ番組のオープニングとエンディングに出現があり，CV で二つに分けたデータセット両方にほぼ同一のショットが含まれてしまったため，重み

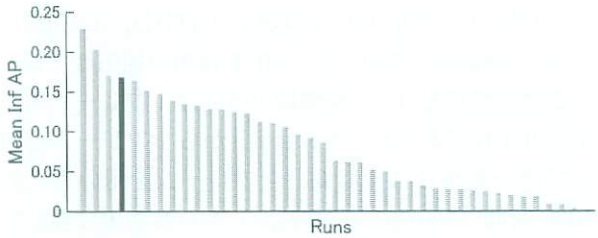

図 7　TRECVID2009 における他手法との Mean Inf AP の比較
Fig. 7　Comparison of Mean Inf AP of the proposed method with the results in TRECVID2009.

の推定がうまくできていない．テストデータにも一部，同一ショットの出現があるため，結果の AP は高くなっているが汎用性を欠いている可能性がある．これは TRECVID のデータセットの問題点であり，評価実験にテレビ番組などの実データを用いる場合，あらかじめ同一のショットのラベル付けを行い，一つのショットとして扱うなどの処理が必要である．

5.3.9　他手法との比較

参考のため，TRECVID2009 における他手法との性能比較を図 7 に示す．この評価は TRECVID2009 development data（100 時間）を学習データ，TRECVID2009 test data（280 時間）をテストデータとした Mean Inf AP で行われている(注1)．図 7 には，TRECVID2009 において評価が行われた 41 個の手法に対する Mean Inf AP を示した．黒く示した部分が SIFT-SVM，SIFT_{hes}-SVM，MFCC-LR の三つを組み合わせた際の結果 (Mean Inf AP = 0.168) である．結果の最高値は 0.228，中央値は 0.063 であり，タスクの難しさがうかがえる．TRECVID2009 のデータでは，実際のテレビ番組を対象に高次特徴検出を行うため，他のデータセットに比べて難易度が高く，大規模なデータに対する頑健性が重視されている．

性能の高い手法では，複数の特徴（色情報を考慮した SIFT 特徴など）及び複数の局所領域を組み合わせたソフトクラスタリング BoW が主流であった．本研究における画像特徴には，2 種類の局所領域による SIFT 特徴（輝度情報のみ）を用いているが，更に多くの要素を組み合わせることで認識精度が向上すると考えられる．

5.3.10　計算時間

実験では，東工大のスーパコンピュータ TSUBAME

(注1)：本論文でこれまでに示した結果は TRECVID2009 development data を 2 分割して学習・テストに用いた際のものであるため直接比較はできない．

(演算サーバ：Sun Fire X4600, 2.4 GHz, 32 GByte メモリ（最大）) を用いた．SIFT 特徴の抽出はフレームごとに処理を行い，特徴抽出時間は 1 フレーム当り平均 0.4 秒であった．各ショットから 1 フレームを選んで用いる場合は 4 時間，全フレームを用いると約 1000 時間（100 cpu で 10 時間）を特徴抽出に要している．

UBM の推定では，128 混合までに 24 時間，512 混合までに 100 時間を要した（サンプル数百万，1 cpu）．MAP 適応（supervector の算出）に必要な時間は 1 ショット当り約 10 秒であった．SVM の学習では，カーネル行列の計算を先に行ったため，1 高次特徴当り平均 15 秒で学習が終了した．結果の融合は，2 手法の場合で 1 秒（0.01 刻み），6 手法の場合で 7 分（0.05 刻み）であった．

以上より，計算量の観点では，特徴抽出に多大な時間を要していることが分かる．今後は，連続したフレームから特徴抽出を行う場合，前フレームの情報などを用いた計算量削減が必要となるであろう．

6. むすび

本研究では，SIFT 特徴と MFCC 特徴を用いた映像からの高次特徴検出手法を提案した．検出方法には，話者認識で用いられているゆう度比による検出と GS-SVM による検出の両者を用いた．実験の結果，単独の手法では SIFT 特徴と GS-SVM による検出が最も良く，Mean AP が 0.141 となった．MFCC 特徴との融合では Singing, Female-human-face-closeup, Person-playing-a-musical-instrument の三つの高次特徴に対して大きな精度向上が見られた．融合手法では最終的に Mean AP が 0.173 まで向上した．しかし，手法に対する重みを事後的に決定すると Mean AP は 0.186 となるため，重みの推定方法には工夫の余地がある．

今後の課題としては，クロスバリデーションの方法の改善やマルチカーネル学習による重み係数の最適化が挙げられる．また，GS-SVM における事後確率推定では，SVM の識別関数の値をシグモイドフィッティングすることで確率値を求めているが，厳密にはロジスティック回帰モデルなどの確率モデルを用いる方がよいと考えられる．また，動画像特徴に関しては，色情報や局所特徴の位置・時間情報を有効活用できる手法が必要である．音響特徴では，音声認識結果を用いることで言語情報を加味した高次特徴検出が考えられる．今後は，動画像情報・音情報・言語情報を組み合わせた高次特徴検出手法が重要となるであろう．

謝辞 本研究は科学研究費補助金基盤研究（B）20300063 の援助を受けた．

文　献

[1] 井上中順，斉藤辰彦，篠田浩一，古井貞熙，“SIFT 混合ガウス分布と音響特徴を用いた映像からの高次特徴検出，”信学技報，PRMU2009-106, Nov. 2009.

[2] D.G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol.60, no.2, pp.91–110, Jan. 2004.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded-up robust features,” Proc. ECCV, 2006.

[4] J. Yang, Y.-G. Jiang, A.G. Hauptmann, and C.-W. Ngo, “Evaluating bag-of-visual-words representations in scene classification,” Proc. MIR, pp.197–206, Sept. 2007.

[5] C.-Y. Lin, B.L. Tseng, and J.R. Smith, “Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets,” Proc. NIST TREC-2003 Video Retrieval Evaluation Conference, 2003.

[6] W. Jiang, C. Cotton, S.-F. Chang, D. Ellis, and A.C. Loui, “Short-term audio-visual atoms for generic video concept classification,” Proc. ACM Multimedia, pp.5–14, 2009.

[7] S.-F. Chang, R. Manmatha, and T.-S. Chua, “Combining text and audio-visual features in video indexing,” Proc. ICASSP, vol.5, pp.1005–1008, 2005.

[8] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation,” Proc. ICASSP, vol.1, pp.97–100, 2006.

[9] A.F. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and TRECVid,” Proc. MIR, pp.321–330, Oct. 2006.

[10] A.F. Smeaton, P. Over, and W. Kraaij, “High-level feature detection from video in TRECVid: A 5-year retrospective of achievements,” in Multimedia Content Analysis: Theory and Applications, pp.151–174, Springer US, 2008.

[11] T. Lindeberg, “Scale-space theory: A basic tool for analysing structures at different scales,” J. Applied Statistics, vol.21, no.2, pp.224–270, 1994.

[12] K. Mikolajczyk and C. Schmid, “Indexing based on scale invariant interest points,” Proc. ICCV, pp.525–531, 2001.

[13] K. Mikolajczyk and C. Schmid, “Scale & affine invariant interest point detectors,” Int. J. Comput. Vis., vol.60, no.1, pp.63–86, Jan. 2004.

[14] Y. Peng, Z. Yang, L. Cao, J. Yi, N. Wan, Y. Feng, X. Zhai, E. Shi, and H. Li, “PKU-ICST at TRECVID 2009: High level feature extraction and search,” TRECVID Workshop, 2009.

[15] K.E.A. van de Sande, T. Gevers, and C.G.M. Snoek, "Evaluation of color descriptors for object and scene recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.9, pp.1582–1596, 2010.
[16] J.C. van Gemert, J.M. Geusebroek, C.J. Veenman, and A.W.M. Smeulders, "Kernel codebooks for scene categorization," Proc. ECCV, pp.696–709, 2008.
[17] F. Perronnin, C. Dance, G. Csurka, and M. Bressan, "Adapted vocabularies for generic visual categorization," Proc. ECCV, pp.464–475, 2006.
[18] N. Rasiwasia and N. Vasconcelos, "Scene classification with low-dimensional semantic spaces and weak supervision," Proc. CVPR, pp.1–6, 2008.
[19] C.G.M. Snoek, K.E.A. van de Sande, O. de Rooij, B. Huurnink, J.R.R. Uijlings, M. van Liempt, M. de Rijke, J.M. Geusebroek, Th. Gevers, M. Worring, A.W.M. Smeulders, D.C. Koelma, M. Bugalho, I. Trancoso, F. Yan, M.A. Tahir, K. Mikolajczyk, and J. Kittler, "The MediaMill TRECVID 2009 semantic video search engine," TRECVID Workshop, 2009.
[20] X. Zhou and S.-F. Chang, "SIFT-bag kernel for video event analysis," Proc. ACM Multimedia, pp.229–238, 2008.
[21] J.C. Platt, "Probabilities for SV machines," in Advances in Large Margin Classifiers, pp.61–74, MIT Press, 2000.
[22] S.J. Young, G. Evermann, M.J.F. Gales, D. Kershaw, G. Moore, J.J. Odell, D.G. Ollason, D. Povey, V. Valtchev, and P.C. Woodland, "The HTK Book, version 3.4," 2006.
[23] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," 2001. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm
[24] M.D. Smucker, J. Allan, and B. Carterette, "A comparison of statistical significance tests for information retrieval evaluation," Proc. CIKM, pp.623–632, 2007.
[25] J. Zhang, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," Int. J. Comput. Vis., vol.73, no.2, pp.213–238, June 2007.

付　　録

検出対象である 20 種類の高次特徴とその説明を以下に示す．これらは，映像検索で用いられることを想定して選ばれており，出現頻度は中程度（データセット中で 100 から 500 回程度の出現）を目安としている．また，† の付いた高次特徴はテレビ番組のオープニングとエンディングに出現がある．

1. Classroom：学校におけるクラスルームのシーン．生徒や先生，黒板などが出現している．

2. Chair：4 本足で一人掛けの椅子（椅子が大きく映っているとは限らない）．

3. Infant：赤ん坊が四つん這いで歩いていたり，抱っこされているシーン．

4. Traffic intersection†：交差点で人や乗り物，信号機などが見えるシーン．

5. Doorway：部屋や建物への入り口となる扉が開いているシーン．

6. Airplane_flying：飛行機が飛んでいるシーン．ただし，気球やヘリコプターは除く．

7. Person-playing-a-musical-instrument：楽器を演奏しており，演奏者と楽器両方が見えるもの．多くの場合，音楽が流れている．

8. Bus：バスが映っているシーン．トラックは除く．

9. Person-playing-soccer：サッカーのシーン．

10. Cityscape：都市の風景で水平線と建物の上部が映っているもの．

11. Person-riding-a-bicycle†：自転車に乗って走っている人が映っているシーン．

12. Telephone：様々な種類の電話．ただし，受話器が見えるもの．

13. Person-eating†：食べ物または飲み物を口に入れているシーン．

14. Demonstration_Or_Protest：複数の人が反対運動を外で行っているシーン．

15. Hand†：人間の手が中心となって映っているシーン．

16. People-dancing：人が踊っているシーン．

17. Nighttime：野外で夜のシーン．ただし，照明があたっているスポーツのシーンを除く．

18. Boat_Ship：水上に船が見えるシーン．カヌーや漕ぎ舟を含む．

19. Female-human-face-closeup：女性の顔が画面の半分以上を占めて映っているシーン．画面に映った女性がしゃべっているシーンが多い．

20. Singing：人が歌っているシーン（BGM のみの場合は負例）．

（平成 22 年 4 月 9 日受付，7 月 27 日再受付）

井上　中順

平 21 東工大・工・情報工学卒．現在，同大大学院情報理工学研究科修士課程在学中．

斉藤　辰彦

平 22 東工大・工・情報工学卒．現在，同大大学院総合理工学研究科修士課程在学中．

篠田　浩一　（正員：シニア会員）

昭 62 東大・理・物理卒．平元同大大学院修士課程了．同年日本電気（株）入社．以来，音声・動画像パターン認識，ヒューマンインタフェースの研究に従事．平 9〜10 ルーセントテクノロジー・ベル研究所客員研究員．平 13 東京大学大学院情報理工学系研究科助教授．平 15 東京工業大学大学院情報理工学研究科助教授，国立統計数理研究所客員助教授．現在，東京工業大学大学院情報理工学研究科准教授．平 9 日本音響学会粟屋学術奨励賞，平 10 本会論文賞各受賞．日本音響学会，IEEE，ACM，情報処理学会，人工知能学会各会員．

古井　貞熙　（正員：フェロー）

昭 43 東大・工・計数卒．昭 45 同大大学院修士課程了．同年 NTT 電気通信研究所入社．昭 53〜54 ベル研究所客員研究員．昭 61 NTT 基礎研究所第四研究室長．平元 NTT ヒューマンインタフェース研究所音声情報研究部長．平 3 同研究所古井特別研究室長．平 9 東京工業大学大学院情報理工学研究科計算工学専攻教授．工博．音声認識，話者認識，音声知覚，音声合成などの研究に従事．科学技術庁長官賞，文部科学大臣表彰，紫綬褒章受章．IEEE より ASSP Society Senior Award，Signal Processing Society Distinguished Lecturer，SP Society Award，James L. Flanagan Speech and Audio Processing Award 受賞．ISCA (International Speech Communication Association) Medal 受賞．本会より，米沢賞，論文賞，著述賞，業績賞，功績賞受賞．日本音響学会より，佐藤論文賞など受賞．著書「ディジタル音声処理」，「Digital Speech Processing, Synthesis, and Recognition」，「新音響・音声工学」，「音声情報処理」など．IEEE，米国音響学会 (ASA) 及び ISCA Fellow．日本音響学会及び ISCA 会長，本会和文論文誌 (A) 及び英文論文誌 (D) 編集委員長など歴任．