社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Bag-of-Frames と時空間特徴量を用いた Semantic Indexing Task への取り組み

下田　保志†　野口　顕嗣†　柳井　啓司†

† 電気通信大学大学院　総合情報学専攻

あらまし　本発表では，国際映像処理ワークショップ TRECVID 2010 で行われた概念索引タスク (Semantic Indexing Task) への取り組みの結果について報告する．映像中のショット動画の全フレームから特徴を抽出し，ベクトル量子化された各フレームの特徴量を要素とする Bag-of-Frames を特徴量として利用した．加えて，時空間特徴，顔特徴など 10 種類の特徴量を抽出し，Multiple Kernel Learning(MKL) によって統合する動作認識フレームワークを用い，Airplane，Boat など 130 種類の概念の認識を行った．その代表の 30 種類の概念についてテストした結果，MKL による統合による結果で平均適合率 infAP=0.0478 を得た，これは TRECVID2010 の参加者 30 チームの内，14 位に当たる．また，10 種類の概念 (light category) においては 37 チームの内，12 位となった．

キーワード　動作認識，時空間特徴，MKL

Approach to conbine Bag-of-Frames and spatio temporal features for TRECVID2010 Semantic Indexing Task.

Yasushi SHIMODA†, Akitsugu NOGUCHI†, and Keiji YANAI†

† Department of Informatics,　The University of Electro-Communication

Abstract　In this paper,　we describe our approach and results for the semantics indexing task at TRECVID 2010.　We focused on spatio-temporal feature and Bag-of-Frames approach.　We used Multiple Kernel Learning as a fusion method to combine these various kinds of features.　Since MKL can estimate weights to combine kernels, as a result of the full-category SIN task,　our MKL approach yielded to get infAP=0.0478 and we reached rank 14 (among 30 teams) for the full-category SIN task and rank 12 (among 37 teams) for the light-category SIN task in TRECVID 2010.

Key words　action recognition,　spatio-temporal feature,　MKL

1. はじめに

本研究では，TRECVID 2010 の Semantic Indexing Task(SIN) を取り上げ，テストコレクションとして提供された動画を分割したショットと，そのキーフレームから，特定の概念を含むショットを抽出するタスクに取り組む．本研究では, [1] で用いられた Bag of Frames，時空間特徴に加え，既存のキーフレーム画像から抽出した特徴量を加え，それら複数の特徴量を Multiple Kernel Learning で統合し，認識対象である 130 種類の内，30 種類の概念を認識する full category と，10 種類の概念を認識する light category について，それぞれ平均適合率を測定した．

2. 関連研究

動画認識ワークショップ TRECVID において近年，マルチフレーム認識が注目を集めている．

アムステルダム大学の MediaMill [2] チームは TRECVID 2008 において，キーフレーム周辺の最大 4 枚の追加フレームから複数の視覚情報を取得したマルチフレーム認識を行いトップの成果を収めた．同チームは翌年も Multiple Kernel Learning(MKL) の一種である SR-KDA(Spectral Regression combined with Kernel Disriminant Analysis) などを利用し，TRECVID 2009 でもトップの成果を得た．

TRECVID2009 において東京工業大学 [3] はショットの全てのフレームから SIFT 特徴を抽出し，また音響特徴も抽出して

それぞれをガウス混合分布を用いてモデル化した．

本手法ではこれらの研究を踏まえ，Bag-of-Frames 表現による全フレーム特徴頻度表現を概念認識に用いる．我々のグループでは，動作認識の研究を行っており，独自の時空間特徴を提案している [1]．SIN タスクでは，“Running”，“Swimming” といった動作を伴う概念も含まれているため，本手法では [1] で提案した時空間特徴も利用する．最も本手法と類似している手法として，Liu らの研究が挙げられる [4]．この手法は特徴量として [5] で提案された時空間特徴を，視覚特徴として SIFT 記述子 [6] を利用し，Adaboost に基づき統合する手法を提案している．またこの研究では，Page Rank に基づく重要な特徴の選択を行っている．

本研究では，視覚特徴，時空間特徴を Multiple Kernel Learning で統合し，さらなる精度向上を試みた．

3. 手　　法

本手法の流れを図 1 に示す．TRECVID によって提供されるキーフレームから取得したカラーヒストグラム，顔特徴と [1] の手法で用いた Bag-of-Frames によって動画全体から取得したガボール特徴，動き特徴，そして，時空間特徴の Bag-of-Feature 表現を Multiple Kernel Learning によって統合し認識を行う．

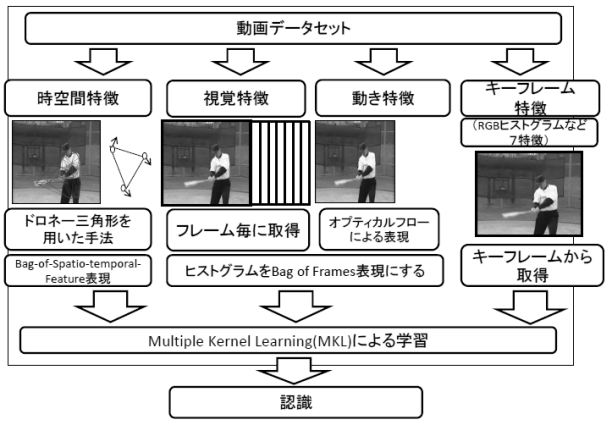

図 1　動画認識の概要

4. 特　　徴

本手法で利用した特徴量を表 1 に示す．動画内の連続した視覚特徴を取得するための時空間特徴における BoF である Bag-of-Spatio-Temporal Feature (BoSTF) 表現，ショット内のフレーム全体の特徴として動き特徴とガボール特徴の Bag-of-Frames 表現，手動アノテーションで用いられるキーフレーム特有の特徴として，6 種類のカラーヒストグラム (HSV，Luv，RGB)(4×3 グリッド，1×1)，顔特徴の 10 種類の特徴量を利用した．

4.1　時空間特徴

手順 4.1 に時空間特徴抽出の流れを示す．

カメラモーションの検出 (Step 1)，視覚特徴の抽出，動き特徴の抽出，最後に特徴の統合である．本項では各段階ごとの説明を行う．

表 1　利用した特徴の概要

特徴	概要	次元
ST_feature	時空間特徴の BoSTF 表現	5000 次元
motion	動き特徴の Bag-of-Frames 表現	3000 次元
Gabor	ガボール特徴の Bag-of-Frames 表現	5000 次元
ch_hsv_globle	HSV カラーヒストグラム (1×1)	64 次元
ch_rgb_globle	RGB カラーヒストグラム (1×1)	64 次元
ch_luv_globle	Luv カラーヒストグラム (1×1)	64 次元
ch_hsv_grid	HSV カラーヒストグラム (4×3)	768 次元
ch_rgb_grid	RGB カラーヒストグラム (4×3)	768 次元
ch_luv_grid	Luv カラーヒストグラム (4×3)	768 次元
face	Haar-Like 顔特徴	1 次元

手順 2 時空間特徴抽出手法の流れ

step1　：カメラモーション検出
step2　：時空間特徴における視覚特徴抽出
step3　：時空間特徴における動き特徴抽出
step4　：視覚特徴ベクトルと動き特徴ベクトルの結合

(Step 1) カメラモーション検出部

一般に，低解像度の動画のみからズームやパンなどの正確なカメラモーションを取得するのは非常に難しい問題である．Liu らはカメラモーションを検出した場合そのフレームを破棄することで対応している [4]．本研究でも同様に、カメラモーションが検出されたフレームからは時空間特徴を取得しない。図 2 のように，グリッド上に Lucas-Kanade アルゴリズム [7] に基づいて，動き特徴を計算する．その中で，動きのあった領域が一定以上であった場合カメラモーションの検出と見なす．

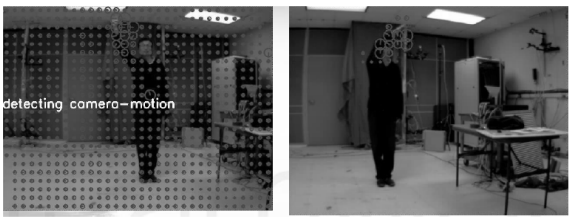

図 2　カメラモーション検出例 (左) と，カメラモーションが検出されない例 (右)

(Step 2) 時空間特徴における視覚特徴抽出部

図 3 は視覚特徴の抽出の様子を示している．図 3 の流れを順番に見ていくと，

(1) フレーム画像から SURF を抽出する．本研究では動画としての特徴を抽出したいので，動きがない点は時空間特徴として適していない．

(2) それぞれの特徴点の動きを計算し，動きがなかった点を削除する．

(3) 最後に残った点について Delaunay 三角分割を行う．視覚特徴には三角形の頂点を構成する三点の SURF 記述子を使

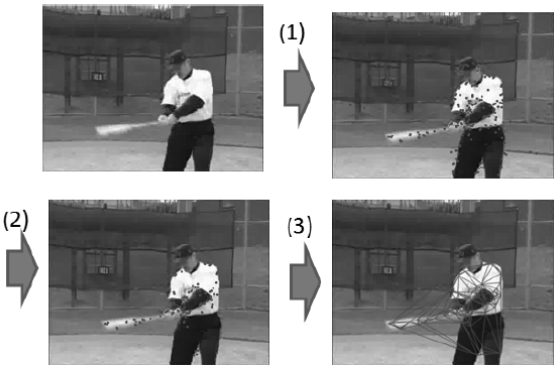

図 3　時空間特徴の抽出の様子

用する．それぞれの三角形の頂点の点を SURF 記述子のスケール毎に整列させる．SURF の次元数は 64 次元なので，結果として視覚特徴は $64 \times 3 = 192$ 次元で表現される．

(Step 3) 時空間特徴における動き特徴抽出部

動き特徴として，本手法では三角形の頂点を構成する，それぞれの点の動きと，三角形の面積の変動を特徴化する．動き特徴は，5 フレーム毎にフレームを M 区間に分割し，それぞれの区間で Lucas-Kanade 法によって，特徴点のオプティカルフローを計算する．各区間の動き情報は x^+, x^-, y^+, y^-, 及び動きなしの 5 次元で表現される．x^+ は x 成分の正方向の，x^- は x 成分の負方向の動き特徴を示す．本手法では，M=5 に設定しているため，各頂点の動き特徴は $(M-1) \times 5$ の 20 次元で，三角形の面積変化は 5 次元で表現される．よって，動き特徴は $20 \times 3 + 5 = 65$ 次元となる．

この手法で得られた動き特徴は回転に関して敏感という問題がある．そのため，SURF 特徴で得られる dominant rotation を利用し，オプティカルフローを回転させる．

(Step 4) 視覚特徴ベクトルと動き特徴ベクトルの結合

時空間特徴は Step 2，Step 3 の処理で抽出した視覚特徴ベクトルと動き特徴ベクトルを単純結合することで $192 + 65 = 257$ 次元の特徴ベクトルで表される．

4.2　Bag of Spatio-Temporal Feature(BoSTF)

本手法では，ショットから取得した時空間特徴の出現頻度のヒストグラムを動画の特徴表現として利用する．Bag of Spatio-Temporal Feature(BoSTF) は画像の表現手法における Bag of Features(BoF) 表現を時空間特徴に適用したものである．

BoF は画像を局所特徴の集合と見なし，座標情報を無視して局所特徴をベクトル量子化し，visual word と呼ばれる代表ベクトルを生成する．

visual word の集合は codebook と呼ばれる．抽出された特徴は codebook の中でも最も近い visual word に投票される．その出現頻度を正規化した上でヒストグラムとして画像を表現する手法である．

BoSTF では局所特徴の代わりに時空間特徴を利用してベクトル量子化を行い，時空間特徴を 5000 次元の BoSTF 表現として利用した．

4.3　motion:動き特徴

時空間特徴は，動き特徴も内包しているが，それだけでは局所的な点の動きしか表現できない．よって本研究ではフレーム全体から動き特徴を抽出する．この動き特徴は，全体的な動きを表現出来るので，時空間特徴点の局所点の動き特徴とは，異なる識別能力が期待できる動き特徴として，本研究ではグリッド点におけるオプティカルフローを利用する．それぞれの検出されたフローは，8 方向，7 段階の大きさからなるヒストグラムに投票され，合計 56 次元の特徴ベクトルとして表される．図 4 は実際に抽出された動き特徴を示している．

図 4　抽出された動き特徴

4.4　Gabor:ガボール特徴

ガボール特徴は画像を 20×20 に分割し各領域から 6 方向 4 周期の 24 次元ベクトルを抽出する．領域ごとのベクトルを結合することによって一つの特徴とする．よって特徴の次元数は各領域のベクトル 24 次元 ×400 の合計 9600 次元で表される．

4.5　Bag of Frames

Bag of Frames は，フレーム一枚を一つの局所特徴と見なすことで，その特徴の出現頻度によって動画を表現する手法である．ガボール特徴の場合，全てのフレームを利用することが可能だが，動き特徴の場合はカメラモーションに対する処理が必要となる．そこで時空間特徴の (Step 1) と同様に，カメラモーションを検出した場合、そのフレームの特徴は無視することで対応した．

本手法では動き特徴を 3000 次元，ガボール特徴を 5000 次元の Bag-of-Frames 表現として利用した．

4.6　キーフレームから取得する特徴

TRECVID で提供されるキーフレームは参加者によるテスト画像の手動アノテーションの際にも用いられるため，重要である．本手法では視覚特徴として 6 種類のカラーヒストグラムと顔特徴を抽出した．

4.6.1　カラーヒストグラム

カラーヒストグラムは RGB，HSV，Luv 色空間を利用した．各色 4 つのビンについて，(64 次元) キーフレーム画像を 4×3 グリッド分割した場合と，キーフレーム画像全体から取得した場合の 2 種類のヒストグラムを作成し，各色空間についてそれぞれ 768 次元と 64 次元の特徴ベクトルとして利用した．

4.6.2　顔　特　徴

顔特徴はキーフレーム画像から．Haar-Like アルゴリズム [8]

を使い抽出した．これは，“Asian People” 等の人物に関係する概念の認識精度向上に用いるため導入した．検出した顔の個数 1 次元の特徴ベクトルとして利用した．

5. 特徴統合

認識する概念によって重要な特徴は異なってくる．概念を認識するためには，各特徴ごとに適切な重みパラメータを与える必要がある．本手法では Multiple Kernel Learning によって特徴の重みを自動で推定し，概念認識に利用した．

5.1 Multiple Kernel Learning

本研究では動作認識を行うために Multiple Kernel Learning(MKL) を利用する．これは複数のサブカーネルを線形結合することで，新たな最適なカーネルを求める手法で，統合カーネルは以下の式で求められる．

$$K_{combined}(\mathbf{x},\mathbf{x}') = \sum_{j=1}^{K} \beta_j k_j(\mathbf{x},\mathbf{x}')$$
$$\text{with} \;\; \beta_j \geq 0, \;\; \sum_{j=1}^{K} \beta_j = 1. \qquad (1)$$

各サブカーネルの重み β_j の設定によって，統合された新たなカーネルの出力は変化する．そのためこの最適な重みをどのように求めるかが問題となり，この問題は MKL 問題と呼ばれている [9]．本手法では，Sonnenburg [10] らによって提案されたアルゴリズムを利用した．

6. 実験結果

Multiple Kernel Learning(MKL) によって特徴の重みを自動で推定した場合 (UEC_MKL) と，単純に特徴量を接合して SVM で学習した場合 (UEC_AVG) で実験を行った．その結果，Full category，light category の平均適合率は表 2，3 のようになった．各参加者グループの中では，UEC_MKL の結果は full category において 30 チーム中 14 位，light category において 37 チーム中 12 位であった．

表 2　TRECVID 2010 Semantic indexing task の結果 (full category)

Category	infAP	Team rank
UEC_MKL full category(30 概念)	0.0478	14/30
UEC_AVG full category	0.0117	
中央値 full category	0.0333	
最高値 full category(MM_CaptainSlow)	0.0900	

表 3　TRECVID 2010 Semantic indexing task の結果 (light category)

Category	infAP	Team rank
UEC_MKL light category(10 概念)	0.0393	12/37
UEC_AVG light category	0.0077	
中央値 light category	0.0180	
最高値 light category(REGIM_5_1)	0.1034	

また，UEC_MKL の Full category における概念ごとの適合率を図 10，実行結果の例を図 7，8，9 に示す．本手法は “Dancing” や “Nighttime”，“Mountain” といった風景の認識で特に良い結果を得た．これは，“Dancing” においては動作の概念であることと，また風景などの概念ではカメラモーションが余りないので，特徴点の追跡を行う時空間特徴が有効に働いたためだと考えられる．一方，路上で見かけるような”Bicycling” や”Running” などは撮影視点の移動によるカメラモーションやズーム，パンが頻繁に発生するため，時空間特徴が出力されず性能があまり上がらなかったと考えられる．

図 11 に示す MKL の重み付けでは，時空間特徴などの高度な特徴が余り大きな重みを示さなかった．これは，カメラモーションのある動画からは時空間特徴や Bag-of-Frames を取得しないため，MKL の最適化の過程で全動画から安定して特徴ベクトルの得られるキーフレーム特徴に一定の重みが与えられたためだと思われる．

7. まとめと今後の課題

本手法は全フレーム特徴頻度表現 Bag-of-Frames を利用し，複数の特徴を MKL で重みづけすることで，ほぼすべての概念で中央値より良い精度を得ることができた．しかし，本手法で用いた特徴量は画像あるいは映像全体から特徴を抽出しているため，複数の物体や概念が同時に写るような映像に対して，十分な検出ができない問題がある．このような概念の検出のための手法の研究が今後必要であると思われる。

文　献

[1] A. Noguchi and K. Yanai: “A surf-based spatio-temporal feature for feature-fusion-based action recognition”, Proc. of ECCV WS on Human Motion: Understanding, Modeling, Capture and Animation (2010).

[2] C. Snoek, K. van de Sande, O. de Rooij, et al.: “The mediamill trecvid 2008 semantic video search engine”, Proc. of TRECVID Workshop (2008).

[3] N. Inoue, S. Hao, T. Saito, K. Shinoda, I. Kim and C. Lee: “Titgt at trecvid 2009 workshop”, Proc. of TRECVID Workshop (2009).

[4] J. Liu, J. Luo and M. Shah: “Recognizing realistic action from videos”, Proc.of IEEE Computer Vision and Pattern Recognition, pp. 1–8 (2009).

[5] P. Dollar, G. Cottrell and S. Belongie: “Behavior recognition via sparse spatio-temporal features.”, Proc. of Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 (2005).

[6] D. Lowe: “Distinctive image features from scale-invariant keypoints.”, International Journal of Computer Vision, pp. 91–110 (2004).

[7] B. Lucas and T. Kanade: “An iterative image registration technique with an application to stereo vision.”, Proc. of International Joint Conference on Artificial Intelligence, pp. 674–679 (1981).

[8] P. Viola and M. Jones: “Rapid object detection using a boosted cascade of simple features”, Proc. IEEE Computer Vision and Pattern Recognition, Vol. 1, pp. 511–518 (2001).

[9] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui and M. I. Jordan: “Learning the kernel matrix with semidefinite programming”, Journal of Machine Learning Research, **5**, pp. 27–72 (2004).

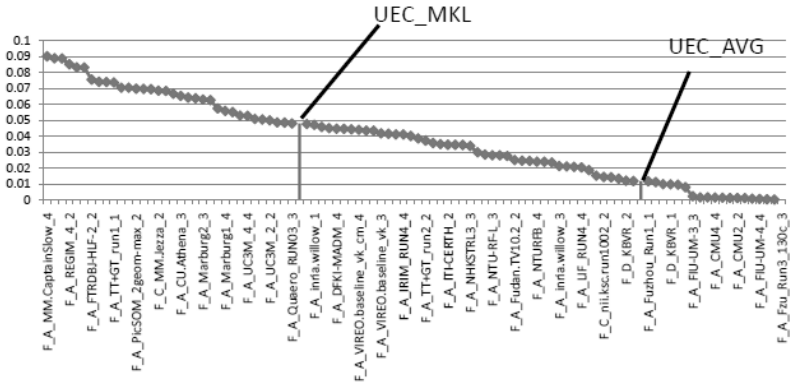

図 5　Full version 全手法の AP 値 (86 手法中，UEC_MKL(32 位) UEC_AVG(70 位))

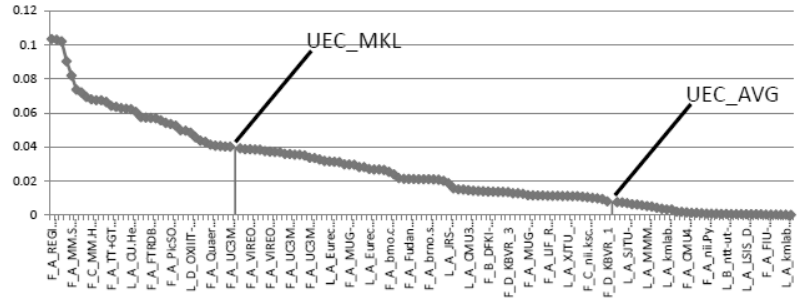

図 6　Light version 全手法の AP 値 (128 手法中，UEC_MKL(31 位) UEC_AVG(94 位))

図 7　Dancing の結果 (UEC_MKL)

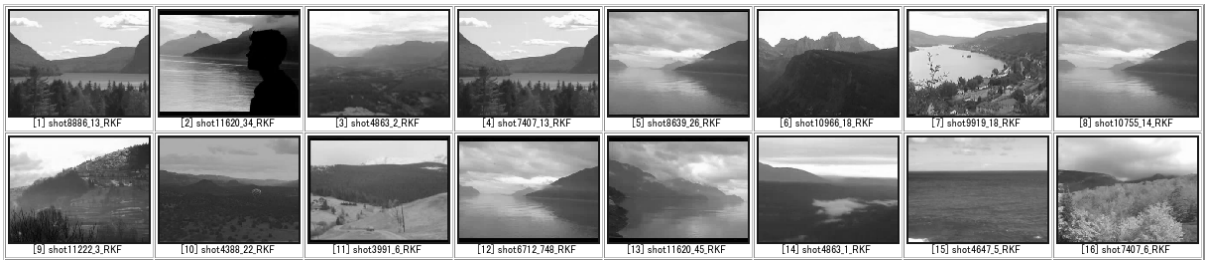

図 8　Mountain の結果 (UEC_MKL)

図 9　Car の結果 (UEC_MKL)

[10] S. Sonnenburg, G. Rätsch, C. Schäfer and B. Schölkopf: “Large scale multiple kernel learning”, Journal of Machine Learning Research, **7**, pp. 1531–1565 (2006).

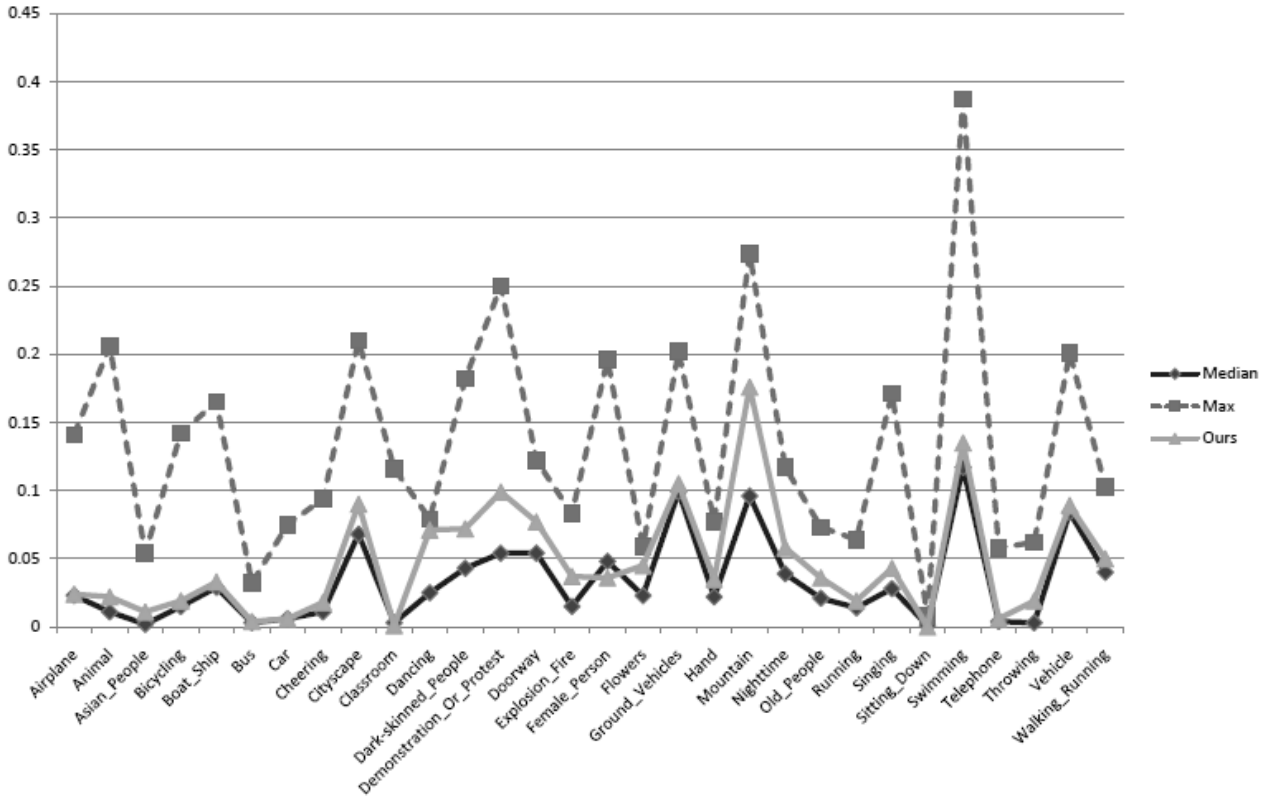

図 10　full version の各概念の認識精度

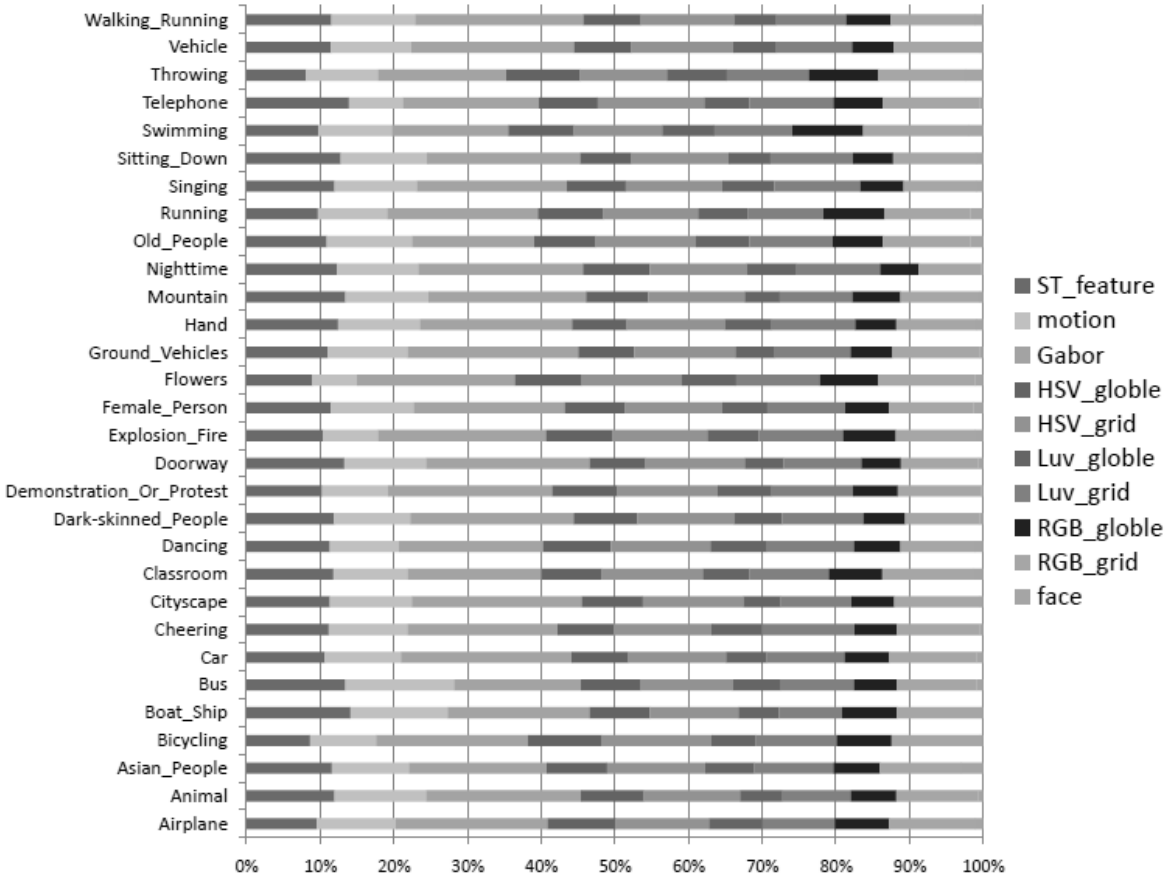

図 11　full version の認識における MKL の特徴毎の重み