H-012

SIFT 特徴量の共起を用いた一般物体認識手法に関する基礎研究
An experitmental study about a generic object recognition based on co-occurrence tendency of SIFT features

下地竜雄馬 † 當間愛晃 ‡ 赤嶺有平 ‡ 山田孝治 ‡ 遠藤聡志 ‡
Ryouma Shimoji Naruaki Toma Yuhei Akamine Koji Yamada Satoshi Endo

1. はじめに

高速データ通信や大容量記憶装置の普及に伴い、画像・映像データ量は爆発的に増加している。そのため、画像・映像情報そのものの意味に即した認識・検索技術が必要とされている。しかし、画像・映像から得られる情報では意味そのものに直接結びつかないため、意味に即した認識・検索は非常に困難だとされている。この問題をセマンティックギャップと呼び、この問題を超えるための試みとして一般物体認識の研究が盛んに行われている [1]。

一般物体認識とは、制約のない実世界シーン画像に対して計算機がその中に含まれる物体を一般的な名称で認識することである。一般物体認識における課題として、特定の物体検出 (Localization) 問題と画像分類 (Categorization) 問題の 2 つが重要な問題とされている。Localization 問題とは、対象とするクラスが画像中のどこにあるのかを認識することであり、Categorization 問題とは、その画像がどのクラスに属するかを認識することである。これらの問題が一般物体認識における重要な課題とされ、一般物体認識を行うためにはこれらの問題を解決する必要がある。

近年では、一般物体認識の手法として、SIFT(Scale-Invariant Feature Transform) 特徴量 [2] を用いた BoF(Bag-of-Features)[3] による認識手法が注目を浴びている。上記で述べている Categoraization 問題として優秀な認識精度をだしたものの、画像中に複数の物体が存在しているようなテスト画像では極端に精度が落ちてしまう。その原因として、BoF はひとつの物体を表現する手法として優秀であるが、画像そのものを BoF を用いてベクトル化することで複数の物体の特徴を含んだ特徴量となってしまうことにより精度が低下していると考えられる。画像中に複数物体が含まれる場合、Localization 問題として画像中にある物体を検出し、それぞれに対して認識するための技術が必要とされる。そのため、本研究では特定物体認識において優秀とされ、BoF にも用いられている SIFT 特徴量から各クラスでの共起情報を求め、それを用いた認識を行う。

2.BoF(Bag-of-Features)

BoF とは、文書分類手法である Bag-of-Words を画像に適応させたものであり、画像を局所特徴の集合とみなすことで局所特徴の位置を考慮せず出現頻度により画像を表現し、その認識を行う。

BoF における処理は、学習と認識に分類される。画像を局所特徴量の集合とみなす BoF の学習では、まず全学習用画像から回転・スケール・照明変化にロバストな SIFT 特徴量の抽出を行う。SIFT 特徴量とは、画像から検出された各キーポイントに対しオリエンテーションを求め、それに合わせて方向の正規化を行ない、各々の持つスケールに対して得られた勾配方向ヒストグラムを正規化することで回転・スケール・照明変化にロバストな特徴量となる。その後、得られた全ての SIFT 特徴量を特徴空間において k-means クラスタリングを行うことにより visual words の生成を行う。各画像から得られた SIFT 特徴量を、生成した visual words から最近傍の visual word に割り当てることでベクトル量子化を行い、visual words 次元のヒストグラムにより表現される。画像毎に抽出される特徴量は異なるため、特徴量の数により正規化し、SVM による学習を行う。

認識では、テスト用画像から SIFT 特徴量を抽出し、それらを学習の際に生成された visual words から最近傍の visual word に割り当てることでベクトル量子化ヒストグラムを求める。その後、正規化を行い、学習により得られた SVM の学習モデルでの認識を行う。

3. 提案手法

本研究では、一般物体認識手法において SIFT 特徴量の共起を用いた認識を行う。学習用画像から SIFT 特徴量を用いた共起を取得することで各クラスに頻出しやすい共起情報を構築する。認識では、テスト画像からも同様に SIFT 特徴量から共起を求め、それらを得られた共起情報を基に画像中に含まれる物体を予測するモデルを提案する。

3.1. 共起情報の取得

学習用画像から SIFT 特徴量の抽出を行う。その際、得られた SIFT 特徴量に対してクラスのラベリングを行う。次に、得られた SIFT 特徴量から visual words を生成し、その後 BoF と同様に各 SIFT 特徴量に対して得られた visual words により最近傍の visual word を割り当てる。visual words は k-means クラスタリングにより生成され、クラスタの総数を K とする。共起情報は画像から得られる SIFT 特徴量において直近の SIFT 特徴量との組み合わせを共起情報として蓄積させる。図 1 に共起情報取得方法を示す。

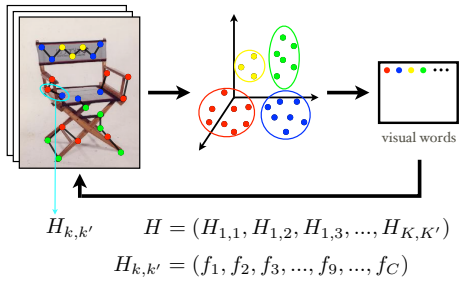

図 1: 共起情報の取得

†琉球大学大学院理工学研究科情報工学専攻
‡琉球大学工学部情報工学科

共起情報とは visual words($k = 1, 2, 3, ..., K$) に割り当てられた SIFT 特徴量の組み合わせ $H_{k,k'}$ とする。共起 $H_{k,k'}$ においてクラス c($c = 1, 2, 3, ..., C$) に出現する頻度を f_c とし、共起情報 $H_{k,k'} = (f_1, f_2, f_3, ..., f_C)(\sum_{c=1}^{C} f_c = 1)$ を求める。ここで、C はクラスの総数であり、$H_{k,k'}$ は k クラスタと k' クラスタとの共起とする。

3.2. 認識

認識では、まず学習と同様に画像から SIFT 特徴量を抽出し、学習過程により生成された visual words を基に各 SIFT 特徴量を最近傍の visual word に割り当てる。その後、各 SIFT 特徴量で割り当てられた visual words での共起を取り、学習過程で得られた共起情報を用いて画像中に含まれている物体の出現頻度を求める。本研究では、出現頻度を求める過程で 2 つのモデルで検証した。図 2 に Model1, 2 の構図を示す。

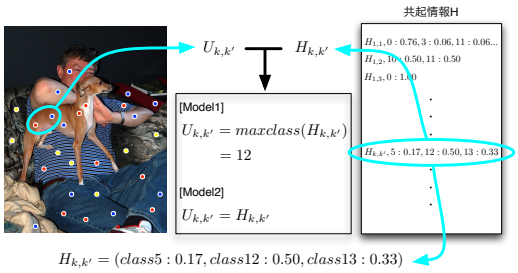

図 2: Model1, Model2

- Model1

  未知画像から検出された共起 $U_{k,k'}$ を、学習過程により得られた共起情報を基にラベリングを行う。ラベリングには共起情報 $H_{k,k'}$ で最も出現頻度の高いクラスを割り当てる。その後、ラベリングされた共起群から各クラスに対して出現頻度 $w = (w_1, w_2, w_3, ..., w_C)$ を求める。

- Model2

  未知画像から得られた特徴量の共起を基に 2 次元ヒストグラム $h = (h_{1,1}, h_{1,2}, h_{1,3}, ..., h_{K,K'})$ を算出する。画像を共起のヒストグラムで表現し、各共起の出現頻度から以下の式により各クラスの出現頻度 $w = (w_1, w_2, w_3, ..., w_C)$ を求める。ここで、f_c は検出された共起 $U_{k,k'}$ におけるクラス c の値とする。

$$w_c = \sum_{k=1}^{K} \sum_{k'=k}^{K} h_{k,k'} * f_c \quad (1)$$

その後、Model1, 2 で得られた出現頻度から最も高いクラスが画像中に含まれているかどうかを確認する。

4. 実験

4.1. 実験方法

実験では、PASCAL[§]データセットを用いて認識・検証を行う。共起情報の取得には、visual words に割り当てられた SIFT 特徴量の組み合わせを使用している。k-means を用いて生成する visual words については、クラスタ数を 500 とする。その際、SIFT 特徴量との距離でソートし、上位いくつまで共起情報として用いいるかを rank1∼5 で実験を行う。また、得られた共起情報とは、k クラスタと k' クラスタとの共起 $H_{k,k'}$ に各クラスでの出現頻度 $H_{k,k'} = (f_1, f_2, f_3, ..., f_C)$ としている。しかし、各共起において出現したクラス数は異なり、1 つのクラスでしか出現しなかった共起から複数のクラスで出現した共起と様々である。そのため、共起情報の使用範囲を range として、1 つのクラスしか出現しなかった共起を unique、2 クラスまでの共起を double、3 クラスまでの共起を triple、全ての共起情報を使用する場合を all とした 4 パターンでの実験を行う。

検証方法として以下の 3 つを行う。

- 既存手法である、BoF による認識率との比較検証を行う。
- 認識率を調べ、各パラメータによりどのような傾向が見られるかを検証する。
- 共起による SIFT 特徴量の認識率を確認し有用な情報が取得できているかを検証する。

4.2. データセット

実験では、PASCAL によるデータセットを使用する。データセットには PASCAL に用意されている Main データセットと Segmentation データセットの積となる画像群を使用し、画像数は学習用画像が 650 枚、テスト用画像が 637 枚の合計 1287 枚を用いて実験を行う。PASCAL から使用したデータセットは表 1 のようになっており、各クラスで画像数が統一されておらず、画像中に含まれる物体数もばらばらである。

表 1: PASCAL

num	Class	train	val
1	aeroplane	37	39
2	bicycle	31	34
3	bird	42	56
4	boat	39	34
5	bottle	33	40
6	bus	29	33
7	car	57	49
8	cat	63	62
9	chair	58	43
10	cow	21	29
11	diningtable	28	25
12	dog	61	51
13	horse	29	33
14	motorbike	34	30
15	person	128	132
16	pottedplant	35	35
17	sheep	27	25
18	sofa	35	35
19	train	32	35
20	tvmonitor	33	30

4.3. 実験結果

まず BoF による認識率と本研究の提案手法における認識率との比較結果を表 2 に示す。

[§]PASCAL2 : Visual Object Classes Challenge 2010 http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/

表 2: 全クラスを平均した実験結果 (%)

	train	val
BoF による認識手法	41.23	23.23
Model1 (rank1, all)	96.62	13.50
Model2 (rank1, all)	92.31	16.64
Model1 (rank5, all)	34.46	7.69
Model2 (rank5, all)	72.92	21.66

表 2 には本研究での Model1, Model2 の中でも未知画像である val で全体的に精度が高かったパラメータ all での rank1, rank5 を載せている。この結果から分かるように現段階では BoF による認識手法を上回る結果は得られていない。しかし BoF を用いた SVM による学習では、学習に使用した画像をテストした場合においても非常に悪い結果となっている。それに比べ、提案手法の Model2(rank5, all) では未知画像においては劣っているものの、共起情報の取得に使用した学習画像に対しては既存の手法を上回っていることがわかる。

本実験では PASCAL のような画像中に複数の物体が含まれるデータセットを用いているため、画像から得られる特徴量を各クラスに分けた上で、BoF によるベクトル量子化ヒストグラムを生成している。そのため、各物体から得られる特徴量は画像中から得られる特徴量全体の 4 割程度となっていることも加わり、SVM による学習が困難だと考えられる。

提案手法では、学習画像から得られた共起情報を基に出現頻度を求めるという統計的手法を用いているため、共起情報の取得に使用した学習画像に対しては安定した認識精度を出すことができている。しかし、未知画像においては未だ認識精度は低いままである。

次により詳しく結果を確認するため、提案手法における Model1, Model2、また各パラメータによる認識精度を図 3、図 5 に示す。

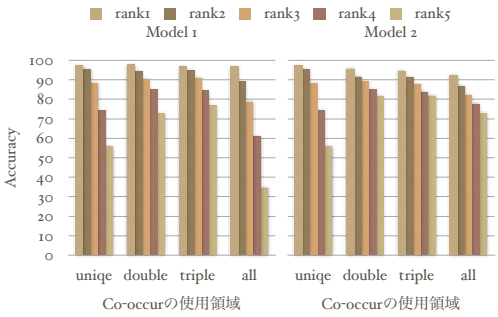

図 3: Accuracy(train)

図 3 が共起情報を取得する際に使用した学習画像の認識精度である。この結果から Model1, Model2 共に高い認識精度が確認できる。Model1 では、共起情報に対してユニークなラベリングを行っているため、Model2 と比較して rank1~5 と情報量が増えるにつれて誤差が大きくなっていることが分かる。また、全体を通して共起情報を使用する range を狭めることでより各クラスで現れやすい共起情報を得られているが、unique, double, triple と徐々に各 rank での揺れ幅が短くなっている。ここで図 4 に各 rank で得られた共起情報から、共起毎に出現したクラス数をカウントしたグラフを示す。

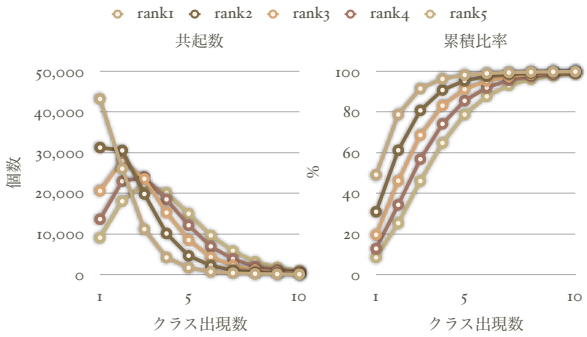

図 4: 左：共起数, 右：累積比率

図 4 右の共起数を示すグラフは、rank1 では綺麗な曲線を描いているが、rank が増える、つまり情報量が増えるにつれてクラス出現数 1, 2, 3 と共起数は減少している。これは、rank により情報量を増加することで、未検出の共起数増加量よりも同じ共起でのクラス出現数の増加量が大きいことがわかる。また、図 4 左の累積比率では、rank1~5 では得られた共起情報数だけでなく、クラス出現数 1, 2, 3 では全体共起数に占める割合が大幅に異なることが分かる。ここでもう一度図 3 を確認してみると、rank1~5 では右肩下がりであり、range 全体でも右肩下がりとなっていることがわかる。これは rank1 が直近の SIFT 特徴量との共起であるのに対し、rank2, 3 と徐々に距離が離れるにつれ関連性が低くなっているためだと考えられ、また range は使用する共起情報を絞ることで重要度の高い共起のみを扱うといえる。この結果から、共起情報を取得する際に使用した学習画像を認識させた場合、rank, range 共に幅を狭めることで各クラスで特徴的な共起のみを使用すると認識率が高くなるといえる。

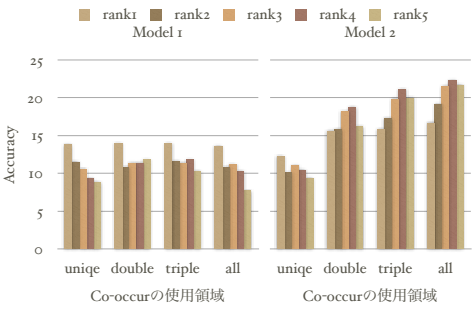

図 5: Accuracy(val)

次に、図 5 の未知画像に対する認識精度を確認する。未知画像に対する認識精度は図 3 とは全く異なる傾向を示しており、rank, range 共に幅を広めた認識精度が高くなっていることが分かる。学習画像を認識する際にはパラメータを絞ることで認識精度を高めていたが、未知画像に対して同様に絞っ

てしまうと扱う情報数そのものを減らすことになるため、出現する共起情報が未検出となってしまうケースが多く見られた。また、未知画像に対して Model1 のように共起情報に対してユニークなラベリングを行ってしまうと、共起情報そのものの揺らぎを無くしてしまうため誤った情報を扱ってしまい、結果認識精度が低くなってしまっている。しかし、比較的認識精度の高いパラメータ (rank5, all) でも２割を超える程度であり、際立った認識精度の上昇は見られなかった。

ここでより詳しく検証を行うため、BoF を用いた SVM による認識手法と本提案手法である Model2 に対して各クラスでの認識率がどのようになっているのかを確認してみた。すると、既存手法と Model2 による認識手法共に認識結果に大幅な偏りが見られた。その原因を検証するため、既存手法、Model2(rank1, all)、Model2(rank5, all) で最も確率が高いと認識されたクラスの画像数での上位３クラスを表３に示し、SIFT 特徴量数、共起数、画像数での上位３クラスを表４に示す。

表 3: 認識手法における出現画像数上位３クラス

	1	2	3
既存手法	person	car	bus
Model2 (rank1, all)	person	cat	dog
Model2 (rank5, all)	person	cat	bus

表 4: SIFT 特徴量数、共起数、画像数の上位３クラス

	1	2	3
SIFT 特徴量数	person	cat	bus
共起数	person	cat	dog
画像数	person	cat	bird

既存手法では「person」「car」「bus」といったクラスに集中して偏りが見られ、Model2 では rank1 で「person」「cat」「dog」、rank5 で上位２つは同じで３位が「bus」となっていた。この表 3, 4 から分かるように、既存手法や提案手法の Model2 においても各クラスでの特徴量数、共起情報、画像数に圧倒的な影響を受けていることが分かる。本実験では、表１のように学習画像の枚数が各クラスによって異なることが特徴的であり、さらに１枚の画像に対して複数の物体が含まれるため物体辺りの特徴量数が画像により大きく変動している。

現在用いている認識手法の Model1, 2 では、主に共起の出現頻度から統計的手法により画像中に含まれる物体の出現確率を求めているため、特徴量数、共起数、画像数の分布に大きく影響を受けてしまっている。これは既存手法でも生じている問題であり、それらから影響を受けないような共起情報の取得方法、または活用方法を改善する必要がある。現在の共起情報の取得には学習用画像から SIFT 特徴量を取得し、visual words のラベリング、その後直近の SIFT 特徴量の共起を取るという過程を経て、全学習用画像から取得した上で各共起に対して相対密度分布を求めている。つまり、ここに特徴量数、共起数の影響を受けている原因があると考える。加えて、取得した共起情報を基に機械学習させることで、より各クラスで特徴的な傾向を学習させることが出来ると考えられる。

5. まとめと今後の課題

本研究では、一般物体認識手法において SIFT 特徴量の共起を用いた認識の実験・検証を行った。SIFT 特徴量の共起を用いた２つのモデルを基に PASCAL データセットを用いて２０種類のクラスの認識実験・検証を行い、その認識精度と有効性とその可能性を示した。

既存手法である BoF を用いた SVM による認識手法での 23.23%という認識精度と比較し、本研究で使用した Model2 での認識精度は最も良かったパラメータで 22.29%と既存の手法を下回る結果となった。しかし、既存手法では学習に使用した画像群に対しても低い認識精度を示したが、Model2 での認識精度では上記のパラメータでは 77.54%の認識精度を示し、最も高いところでは 90%を上回った。また、共起情報の分布や実験から得られた結果により、各クラスに対して有効的な共起情報は取得できているものと考えられる。しかし、実験結果で述べたような PASCAL データセットの特性上、画像中に含まれる複数物体に対しても考慮する必要があるため、共起情報の取得に関しても改善の余地があると考えられる。

本研究における提案モデルは共起情報の有効性を確認するための基礎研究としている。今後の課題として、全学習画像から得られた共起の出現頻度を基に共起情報を構築していたが、各画像から得られる共起での重要性を考慮した取得に改善することでさらに共起情報の有効性を向上させたいと考えている。また、現段階ではどの共起がどのクラスに頻出しやすいという情報を単純な出現頻度で求めているが、それ自体を機械学習により学ばせることで認識精度の向上も図れると考えられる。

参考文献

[1] 柳井啓司：セマンティックギャップを超えて：人工知能学会誌, Vol. 24, No. 5, pp. 691-699, 2009

[2] D. G. Lowe : Distinctive image features from scale-invariant keypoints : International Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110, 2004

[3] G. Csurka, C.R. Dance, L. Fan, and C. Bray : Visual categorization with bags of keypoint : Proc. of European Conference on Computer Vision, pp. 1-22, 2004

[4] Corinna Cortes, Vladimir Vapnik. : Support-vector networks : Machine Learning, 20(3):273-297, 1995

[5] 岡部孝弘：カテゴリの共起を考慮した物体認識：画像の認識・理解シンポジウム, MIRU2008

[6] 永橋知行, 伊原有仁, 藤吉弘亘：画像分類における Bag-of-features による識別に有効な特徴量の傾向：IPSJ SIG Technical Report, Vol.2009-CVIM-169, No.3