2N-01

類似シーン画像を用いた物体検出のフィルタリング

井上直人　古田諒佑　山崎俊彦　相澤 清晴
東京大学

1　はじめに

物体検出においては，近年 CNN [1] を用いた R-CNN [2] やそこから発展した Fast R-CNN [3] が主流である．他手法で得た物体の候補領域を入力として CNN で抽出した特徴量を SVM で学習し，候補領域に対してクラスとスコアを出力する手法である．Fast R-CNN では従来主流の手法であった DPM [4] に対し Average Precision の全クラスでの平均 (mAP) で 30% 以上の精度向上が実現されている．しかし既存の手法では同一画像に対し複数の検出窓を得た時，その出力間の位置関係や大きさの妥当性，そして画像全体の大域的な情報も考慮されておらず，それぞれの窓領域において独立に判定されているのみである．そこで本研究では主にシーンの類似性に注目して抽出した文脈情報を SVM を用いて統合し，物体検出の精度を向上させる手法を提案する．

2　提案手法

提案手法の処理の流れを Fig. 1 に示す．それぞれの要素について，下記で紹介していく．

2.1　文脈情報の抽出とスコアによるフィルタリング

物体検出の学習に使ったデータセットから文脈特徴量 $l_{ap}, l_{rp}, l_{rs}, l_g$ を抽出する．places-CNN [5] により，ある画像 i に対して 205 個のシーン $s_j (j = 1, \cdots, 205)$ への帰属確率 $p(s_j|i)$ が得られる．これとデータセット中の画像における各クラス (c とする) の出現情報を統合することで $p(s_j|c)$ が，さらにベイズの定理より $p(c|s_j) = \frac{p(s_j|c)p(c)}{p(s_j)}$ が得られる．これより画像 i でのクラス c の存在確率 $p(c|i) = \sum_j p(c|s_j)p(s_j|i)$ が得られる．画像 i 中で検出されたクラス c のシーン情報との整合性を示す尤度 l_g を，標準 sigmoid 関数 $\varsigma(x) = \frac{1}{1+e^{-x}}$，データセットの画像枚数 N を用いて $l_g = \varsigma(\frac{p(c|i)}{\frac{1}{N}\sum_i p(c)})$ とする．類似シーン画像に出現する各クラスの位置の分布を混合正規分布で近似することで，検出窓の出現位置の尤度 l_{ap} も得られる．検出窓同士の位置関係の尤度 l_{rp}，検出窓同士の大きさの関係の尤度 l_{rs} はデータセット全体の情報から l_{ap} と同様に算出する．いずれの尤度 l も $0 \leq l \leq 1$ に正規化されており大きいほどより検出が尤もらしいことを意味する．

また検出窓のスコアが 0.05 を下回る領域は全体の候補領域の 9 割程度を占めるが，ここに正しい検出は殆ど含まないため，無条件に棄却する．

2.2　SVM による特徴選択と物体検出の再評価

抽出された検出窓の文脈特徴量及びスコアと正誤のラベルのセットを用いて各クラス毎に，検出窓のスコアと特徴が与えられた時にそれが正しい検出か否かの 2 クラス分類器を非線形 SVM (RBF カーネル) で学習する．テスト時には検出窓から特徴量を同様に抽出して分類器で識別し，得られた decision value の値 $d \in \mathbb{R}$ に対して新しいスコア $s = \frac{1}{1+e^{-2d}} (0 \leq s \leq 1)$ を得る．また，SVM の学習の前段階の処理として特徴選択を行う．教師データをさらに SVM の学習用と評価用に 3:1 で分割し，スコア以外の 4 つの特徴量についてどの特徴量を使うかの計 $2^4 = 16$ 通りに関して，学習用データで学習し評価用データを識別した際の AP が最も高くなる組み合わせを採用して前述したスコア計算を行う．

3　実験

物体検出には Fast R-CNN，データセットには VOC2007 [6](画像数: 9963 枚，クラス数: 20 クラス) を用いた．Fast R-CNN の学習と尤度情報の抽出と再評価のための SVM の学習に VOC2007 のうち trainval (5011 枚) を用い，SVM による検出結果の再評価を VOC2007 のうち test (4962 枚) を用いて行った．Fast R-CNN そのものの出力，スコアのフィルタリング

Filtering Object Detection Results Featuring Statistical Information of Objects in Similar Scenes
Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, Kiyoharu Aizawa
The University of Tokyo

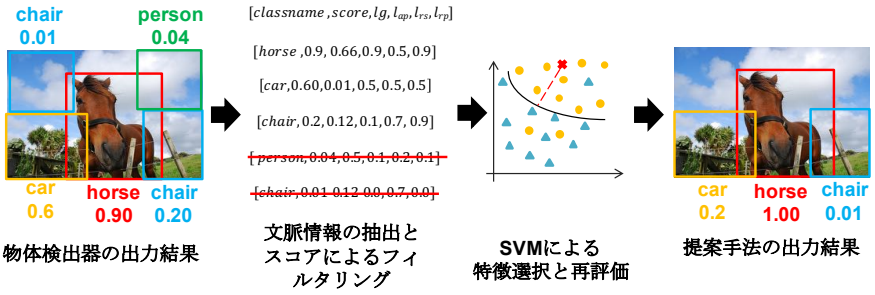

Fig. 1　提案手法の流れ

Table. 1　Average Precision (%) and F1 (%) on VOC 2007 test

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP	F1
Fast R-CNN [3]	73.8	**78.2**	68.7	**54.4**	37.0	**76.4**	78.2	**82.8**	41.4	**71.7**	**67.7**	76.9	**78.1**	**74.0**	66.9	33.0	62.4	**68.9**	74.0	67.9	66.9	3.50
Fast R-CNN($score > 0.05$)	73.8	77.7	68.7	54.4	36.8	75.9	78.2	82.8	41.0	71.7	67.7	76.9	78.1	73.0	66.5	31.6	62.4	67.5	77.8	66.9	66.5	**27.7**
proposed method	**75.0**	77.9	**69.1**	53.5	**37.7**	76.0	**78.4**	82.3	**44.0**	71.4	**67.5**	**77.0**	78.0	73.6	**67.5**	**35.6**	**62.8**	68.6	**74.8**	**68.8**	**67.0**	**27.7**

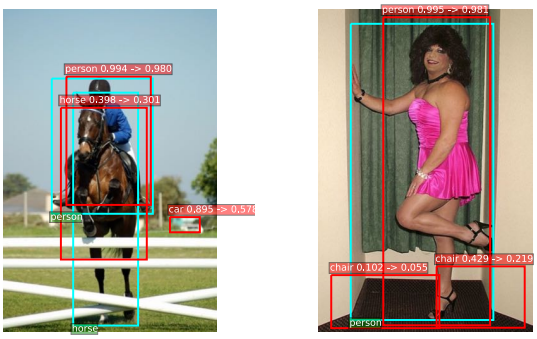

Fig. 2　成功例

($score > 0.05$) 後の出力，再評価後のクラス毎の AP とその平均値 (mAP)， F1 を Table. 1 に示す．

Table. 1 より特に元々の AP の低い，つまり検出が難しいクラスである pottedplant (表では plant と表記)，chair 等のクラスに対して AP で 2% 以上の改善を実現していることがわかる．以上のように AP の改善も果たしながら，提案手法では正解領域の割合が少ない $score \leq 0.05$ の領域について棄却することにより，大幅な候補領域数の削減を果たし F1 では 24% 以上の改善を実現している．実際に成功した例を Fig. 2 に示す．青枠が ground truth の領域，赤枠が物体検出によって得られた候補領域であり，スコアは物体検出そのものによるスコアと SVM による再評価によってつけたスコアの双方が示されている．左の図では car が，右の図では chair が，それぞれ画像の文脈情報にもとづいて person に比べ大幅にスコアが下がっていることがわかる．

4　結論

物体検出において文脈を考慮して検出窓のスコアを再計算することで，物体検出の精度を保ちつつ，F1 を 24% 以上改善させ検出窓のうち False Positive を大幅に削減できることが確認された．本手法は任意の物体検出の手法に対して後処理の形で付加することが出来る．今後は特徴量選択をさらに工夫するとともに，より多くのクラスを含むデータセットに本手法を適用したい．

参考文献

[1] Y. LeCun et al. Backpropagation applied to handwritten zip code recognition. *Neural computation*, Vol. 1, No. 4, 1989.

[2] R. Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[3] R. Girshick. Fast R-CNN. In *ICCV*, 2015.

[4] P. Felzenszwalb et al. Object detection with discriminatively trained part-based models. *TPAMI*, Vol. 32, No. 9, 2010.

[5] B. Zhou et al. Learning deep features for scene recognition using places database. In *NIPS*, 2014.

[6] M. Everingham et al. The pascal visual object classes (voc) challenge. *IJCV*, Vol. 88, No. 2, 2010.

東京大学 工学部 電子情報工学科
〒113-8656 東京都文京区本郷 7-3-1
TEL.03-5841-6761
E-mail: naoto@hal.t.u-tokyo.ac.jp

(謝辞) 本研究の一部は科学研究費補助金 (26540078)，及び一般財団法人テレコム先端技術研究支援センター研究助成 (SCAT) の支援プログラムを受けて行われた．