H-028

簡易な教示データを用いた multiple-instance 型 LVQ による頭部検出

Head detection based on multiple-instance LVQ by using easily annotated data

細井　利憲†　　今岡　仁†　　宮野　博義†　　石寺　永記‡
Toshinori Hosoi　Hitoshi Imaoka　Hiroyoshi Miyano　Eiki Ishidera

1．はじめに

頭部など特定の物体を検出するには，事前に学習された「物体／ 非物体」の 2 クラス識別器で sliding-window 探索すると高い検出精度を得られる．しかし，あらゆる条件下の学習データを事前に用意することは困難であり，学習データの物体と実環境中の物体との見え方が異なる場合には識別精度が低下する．これを解決するために実環境の物体データを集めて追加学習すれば精度を高められるが，実映像中の特定物体の位置・サイズ・回転等の情報を教示する必要がある．そのため，十分なデータを集めようとすると，教示作業の期間と人的コストが膨大になりがちな点が実用上の大きな問題である．

教示作業を軽減するために，映像中のある 1 フレームの物体を人手で正確に教示し，残りのフレームでは自動的にトラッキングされた物体領域を学習データとする枠組みが提案された[1]．個々の物体領域のパッチ $\hat{\mathbf{x}}_i$（インスタンスと呼ぶ）を基に位置・サイズ・回転の摂動を与えた多数のパッチ $\mathbf{x}_{ij}, j = 1, \ldots, N_j$ を生成し，これらをまとめて 1 つのバッグとする（図 1, 4 参照）．トラッキングを継続すると位置ズレが発生し粗いアライメント情報となるが，摂動することで「バッグ中の少なくとも 1 つのインスタンスは正しい」と期待される．このようなバッグ単位の学習には Multiple Instance Learning (MIL)が利用される．筆者らは一般化学習ベクトル量子化（GLVQ）ベースの高速な MIL を提案し，粗いアライメントであっても追加学習の効果を確認できた．しかし，トラッキングに完全に失敗するなど自動教示に大きな誤りがあると，それを摂動しても正しいインスタンスを 1 つも生成できず，追加学習の効果を得られない問題がある．

そこで本稿では，非常に簡易な教示作業「物体中の任意の 1 点のみ入力」を実施することで，教示に大きな誤りがなく MIL に適した物体領域パッチを生成し，学習効果を確実に得られるフレームワークを提案する．簡単に入力できる粗い 1 点から頭部領域を推定し，摂動させたパッチ群を MIL で追加学習する．尚，本稿では検出対象物体を「人体の頭部」とし，人体に関する事前知識も利用する．

2．提案フレームワーク

既に学習済みの頭部検出エンジンが存在し，それを特定の実環境の映像で高精度に動作するよう追加学習する場合を想定する．本提案フレームワークは次の通りである．

1. 粗いフレーム間隔で頭部中の 1 点を手入力する
2. 1 点から頭部領域を推定する
3. 未入力フレームの頭部領域を補間する
4. 頭部領域を摂動させて，バッグを生成する
5. MIL により追加学習する

† 日本電気(株)　情報・メディアプロセッシング研究所
‡ (株)NEC 情報システムズ 先端技術ソリューション事業部

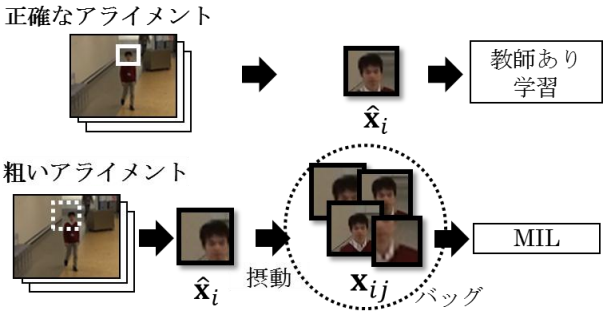

図 1　追加学習の作業の比較

図 2　1 点の入力位置（点線エリア内の任意の 1 点）

2.1　1 点教示と頭部領域推定

まず，人手により頭部の 1 点を教示する．具体的には，頭部の中心付近の任意の 1 点を入力し，図 2 に示すように点が頭部内にあれば許容するものとする．このように要求される位置精度が高くなければ，作業者の集中力や正確な機器操作は不要であり，初心者でも簡単に入力できる．

次に，入力された 1 点から画像上の頭部領域を推定する．事前にカメラの内部・外部パラメータを推定し [3]，世界座標系上の人体の身長と頭部サイズをそれぞれ固定値，地面を平面と仮定すれば，1 点を世界座標系に変換しかつその 1 点を中心とした人体の頭部サイズの球を頭部領域と仮定できる．この球領域を画像座標系に逆変換し外接矩形をとることで，画像上の頭部領域が 1 つ得られる．

次に，未入力フレームの頭部領域を推定済みの頭部から補間する．具体的には，頭部領域の位置と見えに関して世界座標系で前後フレームからトラッキングする．ただし，トラッキングの信頼度が低い場合にはその結果を使わない．提案方式では 1 点の教示が容易なため，教示フレーム間隔を密にし易く，この制約をほぼ受けないようにできる．

以上により推定された頭部領域 $\hat{\mathbf{x}}_i$ を位置・スケール・回転に関して摂動させたインスタンス群 $\mathbf{x}_{ij}, j = 1, \ldots, N_j$ でバッグを生成すれば，「バッグ中の 1 つ以上のインスタンスが正しい」という MIL の条件をほぼ確実に満たせる．

2.2　MIL による追加学習

本フレームワークでは摂動させた多数のパッチを利用するため，学習が高速な GLVQ ベースの MIL を利用する[2]．この手法では，バッグ i内の j番目のインスタンス $\mathbf{x}_{ij}$ が正

しいクラス z_iである確率 p_{ij}を用いて，バッグ i 中の少なくとも 1 つのインスタンスが正しいクラス z_iである確率 q_iを式(1)のように表現し，式(2)で定義される損失 Lを最小化するよう学習される．

$$q_i(z_i|\mathbf{x}_{i1},\dots,\mathbf{x}_{iN_i}) = \begin{cases} 1-\prod_{j=1}^{N_i}(1-p_{ij}(z_i|\mathbf{x}_{ij})) & , if\ Pos. \\ \prod_{j=1}^{N_i} p_{ij}(z_i|\mathbf{x}_{ij}) & , if\ Neg. \end{cases} \quad (1)$$

$$L(q_i) = -\prod_{i=1}^{N} q_i(z_i|\mathbf{x}_{i1},\dots,\mathbf{x}_{iN_i}) \quad (2)$$

3. 頭部検出実験

3.1 実験条件

公開人体画像セット INRIA Person Dataset[4]の頭部を GLVQ により学習したものを追加学習前の頭部検出辞書とする．これに対し，図 3 を現場カメラの想定映像とし，先述の 1 点を教示し，GLVQ により追加学習した場合と，LVQ 型の MIL により追加学習した場合を比較した．ただし，頭部データの追加効果を明確に測るために，現場の非物体データは利用しない．一方，検出率の評価には同一カメラで撮影された 13 シーンを用い，検出結果の正解判定にはほぼ正確に手入力した矩形と比較した．

学習データは 10fps で撮影された計 14,900 フレームの映像で，1 点入力のフレーム間隔を 10 とした．ただし，頭部の移動が直線的であれば入力を省略したため，実際の入力間隔は 10～70 となった．一方，頭部矩形推定のための人体身長は 160cm, 頭部幅を 15cm と仮定した．この設定で得られた実際の頭部パッチを図 4 に示す．図のように頭部位置・スケールのばらつきは比較的大きいが，著しい誤りがあるパッチは確認されなかった．このような各パッチから摂動によって 96 個の学習用パッチを生成した．

識別・検出処理では，頭部パッチサイズは 30×30 画素，特徴量は輝度勾配(512 次元ベクトル)とし，重複検出された候補のマージには Non Maximum Suppression を用いた．

3.2 実験結果

各学習結果について，現場映像からの頭部検出率を図 5 に示す．横軸は Precision，縦軸は Recall であり，凡例の”PublicData(GLVQ)”は追加学習前の場合，”Public + ReadData(GLVQ)”は 1 点入力で生成された現場の学習データを GLVQ で追加学習した場合，”PublicData(GLVQ) + RealData(IL-LVQ)”は同様の学習データを MIL-LVQ で追加学習した場合を表す．この図より，追加学習によって現場映像での検出率が Precision, Recall 両面で改善した．特に，MIL-LVQ では通常の教師あり学習(GLVQ)よりも改善効果が高く，追加学習によって検出漏れが 30～65%削減された．このことから，本実験では 1 点教示された学習データが MIL の条件を十分満たせたといえる．

ところで，文献[2]の実験では通常の教師あり追加学習で識別率が改善しなかったが，本実験では改善した．これは，1 点の手動教示によって学習パッチの大きな誤りが無くなり，アライメントが比較的正確になったためと考えられる．

4. おわりに

頭部検出を特定環境に対応するための追加学習を容易に行えるフレームワークを提案した．簡単に入力できる粗い 1 点のみ手動で教示し，頭部領域を推定し，摂動させたパッチを MIL で学習する．現場カメラ映像に対応させる実験では，頭部の検出漏れを 30～65%削減し検出率を改善できる事を確認した．今後の課題は，作業量の定量評価，省力化，MIL により効果が得られる他の条件の明確化が挙げられる．

参考文献

[1]P. Viola, J.C. Platt, C. Zhallg, “Multiple Instance Boosting for Object Detection”, NIPS2005, pp.1417-1424, 2005

[2] 細井, 宮野, 石寺, “Multiple Instance LVQ による物体検出”, 信学技報, PRMU2011-74, pp. 145-150, 2011

[3] Tsai, Roger Y., “An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision,’’ CVPR1986, pp. 364-374, 1986

[4] N. Dalal, B. Triggs, “Histograms of Oriented Gradients for Human Detection”, CVPR2005, vol. 2, pp.886-893, 2005

図 3　学習用現場映像

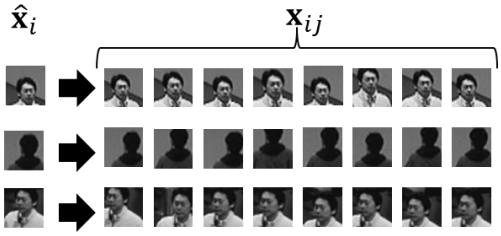

図 4　学習用の頭部領域と摂動の実例

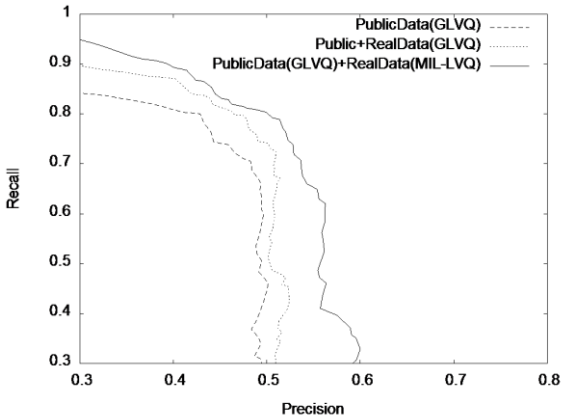

図 5　現場画像からの頭部検出率