H-007

時系列深層学習と一般物体認識ネットワークを用いた物体追跡
Visual Object Tracking by Using Generic Object Recognition and Convolutional LSTM Network

村手翼 [1]　渡辺崇 [1]　山田正生 [2]

[1] 名古屋大学院 情報学研究科
[2] 元名古屋大学院 情報学研究科

1　はじめに

物体追跡とは動画像内に置いて対象物体の存在領域を推定する問題である．実世界にて人間が物体追跡をする際, 人間は視覚情報のみを用いるのではなく, 視覚情報をベースとしたある程度の記憶と予測を用いて物体追跡を行う．本研究ではその人間の物体追跡を模倣するというアイデアのもと, 視覚的情報を用いる物体追跡手法をベースに記憶予測機能を付加し精度向上を目指す．ベースとする物体追跡手法には, 一般物体認識の畳み込みニューラルネットワーク (Convolution Neural Network) を応用したものを用いた．そして, 記憶予測機能は畳み込み層を持った LSTM(Long Short Term Memory) の ConvLSTM を用いて再現した．

2　Fiture Map Selection Tracking

一般物体認識 CNN を用いた物体追跡手法には, Fiture Map Selection Tracking(FMST) [5] を用いる．この手法は学習済み 16 層 VGG Net [2] を用いて, 追跡対象のセグメンテーションを抽出し, 対象の存在領域を推定する手法である．

2.1　特徴マップ選択

対象の推定存在領域に多く活性が発生する特徴マップを選択するために重要度を計算する．目標マップ $\mathbf{M}^l$ を用いて, ROI を CNN に入力して得た第 l 層の特徴マップの c チャンネル目を $\mathbf{F}_c^l$ として, $\mathbf{F}_c^l$ の重要度を次式で定義する．

$$s_c^l = \mathrm{sum}(\mathbf{F}_c^l \circ \mathbf{M}^l) \tag{1}$$

ただし $\circ$ は行列の要素毎の積 (アダマール積), sum は行列の全要素の和を表す．

第 l 層について重要度 s_c^l が上位であるチャンネルの集合 C^l についての特徴マップの和を以下のようにとる．

$$\hat{\mathbf{M}}^l = \sum_{c \in C^l} \mathbf{F}_c^l \tag{2}$$

これを $l \in \{4, 5\}$ について得て, 和を予測マップ $\hat{\mathbf{M}}$ とする．

3　提案手法

3.1　提案手法の概要

FMST の手法で得られる, 視覚的情報の予測マップを $\hat{\mathbf{M}}_\alpha$ とし, 記憶予測機能にて得られる予測マップを, $\hat{\mathbf{M}}_\beta$ とする．そして, その 2 種類の予測マップから次式にて新たな予測マップ $\hat{\mathbf{M}}_\gamma$ を得る．

$$\hat{\mathbf{M}}_\gamma = p\hat{\mathbf{M}}_\alpha + (1-p)\hat{\mathbf{M}}_\beta \tag{3}$$

$\hat{\mathbf{M}}_\gamma$ を最終的な予測マップとして, そこから予測領域を導き出す．基本的には $p = 1.0$ とし, $\hat{\mathbf{M}}_\alpha$ のみで追跡を行う．正規化相互相関を用いて前後フレームでの予測が大きく変化した場合にのみ, $p = 0.8$ として $\hat{\mathbf{M}}_\beta$ を $\hat{\mathbf{M}}_\gamma$ に加味するという形をとる．また, $\hat{\mathbf{M}}_\beta$ を加味する場合は $\hat{\mathbf{M}}_\alpha$ の活性を減衰させて足し合わせる．

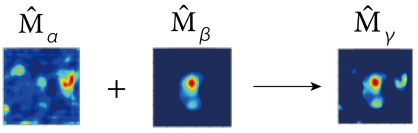

図 1: $\hat{\mathbf{M}}_\gamma$ の生成例

3.2　ConvLSTM と正規化相互相関

時系列データの予測に用いられる LSTM [1] に畳み込み層を持たせ, 動画像の予測を可能にした

ConvLSTM[6] にて記憶予測機能を再現する．過去フレームから対象の存在領域を予測するマップを, 予測マップ $\hat{\mathbf{M}}_\beta$ とする．

$\hat{\mathbf{M}}_\alpha$ のみを用いた追跡にて直前時刻から大きく予測マップが変化した場合, 予測が転移している可能性が高い．変化が大きい場合に $\hat{\mathbf{M}}_\beta$ を $\hat{\mathbf{M}}_\gamma$ に加味する．この変化度を測るために正規化相互相関を用いた．ここでは相関の評価値を類似度 R と呼び, P と Q の類似度 $R(P,Q)$ は以下のように表す．

$$R(P,Q) = \frac{\sum_y^Y \sum_x^X P(x,y)Q(x,y)}{\sqrt{\sum_y \sum_x P(x,y)^2 \sum_y \sum_x Q(x,y)^2}} \quad (4)$$

予測マップ $\hat{\mathbf{M}}_\alpha$ は ROI というフレームの一部分の表現である．フレーム全体のでの位置を考慮するため, フレームと同サイズのゼロ行列に $\hat{\mathbf{M}}_\alpha$ の予測領域 $\mathbf{x}_t$ に該当する部分を埋め込んだマップ $\hat{\mathbf{M}}_\delta$ にて類似度を評価する．

3.3 提案手法の全体

Algorithm 1 提案トラッキング手法

Input: 正解領域 $\mathbf{x}_0$，画像 $\mathbf{I}_t$，$t \in \{0, 1, ..., T\}$
Output: 予測領域 $\mathbf{x}_t$, $t \in \{1, 2, ..., T\}$

$\mathbf{I}_0$ から $\mathbf{x}_0$ 中心の ROI を切り抜き $\mathbf{F}^l$ を得る;
$\mathbf{x}_0$ と $\mathbf{F}^l$ から選択する特徴マップを初期化する;
for $t = 1$ to T **do**
    $\mathbf{I}_t$ から $\mathbf{x}_{t-1}$ 中心の ROI を切り抜き $\mathbf{F}^l$ を得る;
    $\mathbf{F}^l$ の中から特徴マップを選択し $\hat{\mathbf{M}}_\alpha$ を得る;
    式 (3) から $p = 1.0$ として $\hat{\mathbf{M}}_\gamma$ を得る;
    $\hat{\mathbf{M}}_\gamma$ 上から予測領域 $\mathbf{x}_t$ を得る;
    $\hat{\mathbf{M}}_\gamma$ と $\mathbf{x}_t$ を用いて $\hat{\mathbf{M}}_{\delta,t}$ を得る;
    if $t > 2$ **and** $R(\hat{\mathbf{M}}_{\delta,t-1}, \hat{\mathbf{M}}_{\delta,t}) < N$ **then**
        式 (3) から $p = 0.8$ とした $\hat{\mathbf{M}}_\gamma$ を得る;
        $\hat{\mathbf{M}}_\gamma$ 上で予測領域 $\mathbf{x}_t$ を再度決定する;
        $\hat{\mathbf{M}}_\gamma$ と $\mathbf{x}_t$ を用いて $\hat{\mathbf{M}}_{\delta,t}$ を得る;
    end if
    $\mathbf{x}_t$ と $\mathbf{F}^l$ を用いて s^l を更新する;
end for

4 評価実験

Wu ら [4] によるデータセットで提案手法の評価を行なった．Intel i7-4790 3.60GHz CPU と 16GB RAM, NVIDIA GeForce TITAN X GPU を搭載した PC で実験を行なった．FCNT は Wu ら [3] による手法で, こちらも特徴マップ選択を行う手法である．処理速度は

	FMST[5]	Present	FCNT[3]
precision (%)	83.68	**87.05**	85.6
success (%)	57.86	**60.36**	59.9
processing speed (fps)	**42.13**	34.87	∼3

FMST から 20 %ほど低下したが, 予測中心座標の精度 (precision) と予測領域の精度 (success) は他の 2 手法を上回る結果となった．

5 終わりに

視覚的情報と記憶予測情報を用いた人間の物体追跡を模倣した手法を提案した．処理速度はやや低下するものの, 追跡精度の向上は確認することができた．

参考文献

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long short term memory. Technical Report FKI-207-95, Technische Universität München, 1995.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[3] Lijun Wang, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu. Visual tracking with fully convolutional networks. In *ICCV - IEEE International Conference on Computer Vision*, 2015.

[4] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR - IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[5] Masaki YAMADA. 畳み込みニューラルネットワークの特徴マップ選択によるトラッキング. In *The 79th National Convention of IPSJ,1P-08*, 2017.

[6] Xingjian Shi Zhourong Chen Hao Wang Dit-Yan Yeung. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *arXiv:1506.04214 [cs.CV]*, 2015.