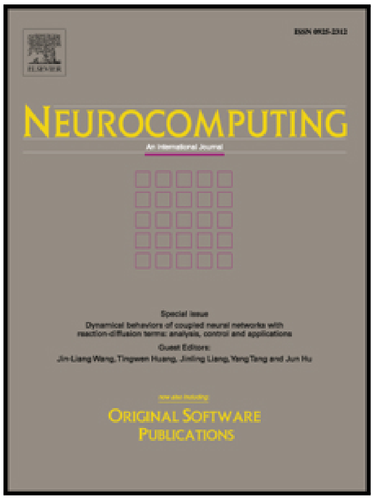
Communicated by Dr Tianzhu Zhang

Accepted Manuscript



A Deep Features Based Generative Model for Visual Tracking

Ping Feng, Chunyan Xu, Zhiqiang Zhao, Fang Liu, Jingjuan Guo, Caihong Yuan, Tianjiang Wang, Kui Duan

PII: S0925-2312(18)30536-8
DOI: 10.1016/j.neucom.2018.05.007
Reference: NEUCOM 19555

To appear in: *Neurocomputing*

Received date: 7 July 2017
Revised date: 16 January 2018
Accepted date: 2 May 2018

Please cite this article as: Ping Feng, Chunyan Xu, Zhiqiang Zhao, Fang Liu, Jingjuan Guo, Caihong Yuan, Tianjiang Wang, Kui Duan, A Deep Features Based Generative Model for Visual Tracking, *Neurocomputing* (2018), doi: 10.1016/j.neucom.2018.05.007

A Deep Features Based Generative Model for Visual Tracking

Ping Feng[a], Chunyan Xu[b], Zhiqiang Zhao[a], Fang Liu[a], Jingjuan Guo[a], Caihong Yuan[a], Tianjiang Wang[a,*], Kui Duan[c]

[a] *School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, 430074, China*
[b] *School of Computer Science and Engineering, Nanjing University of Science and Technology, 210094, China.*
[c] *Huazhong University of Science and Technology school hospital, Wuhan, 430074, China*

Abstract

In this work, we propose a novel visual tracking algorithm based on a framework of generative model. In order to make the algorithm robust to various challenging appearance changes, we adopt the powerful deep features in the description of tracking object appearance. The features are extracted from a Convolutional Neural Network (CNN), which is a modified one based on the VGG-M nets but constructed with fewer convolution layers and sequences exclusively full connection layers. In the pretraining process, we add a special convolution layer called coefficients layer before the full connection layers. In the tracking process after the network being pretrained, we remove the coefficients layer and just update the full connection layers conditionally. To decide the new target's positions, we compute the compositive similarity scores containing three kinds of similarities with different weights. The first kind is similarities between candidates and the target in the first frame, and the second kind is between candidates and tracking results in the last frame. The third kind is related to the important object appearance variations in the tracking process. We design a simple mechanism to produce a collection to record those historical templates when the object appearance changed largely. With similarities between candidates and the historical templates, the drift problem can be alleviated to some extent, because similar historical appearances sometimes appear repeatedly and the recorded historical templates can provide important information. We use the outputs of the convolution part before the full connection layers as features and weight them with the coefficients layer's filter weights to compute all similarities. Finally, candidates with the highest scores will be regarded as new targets in the current frame. The evaluated results on CVPR2013 Online Object Tracking Benchmark show that our algorithm can achieve outstanding performance compared with state-of-the-art trackers.

Keywords: visual tracking, deep features, generative model

1. Introduction

Visual tracking is a fundamental research area in computer vision. As several kinds of important applications (like scene surveillance system, intelligent traffic control system, motion pattern analysis, etc.) are closely related to it, much attention has been drawn to develop robust tracking algorithms. Though great progress has been made in recent years, more effort should be made to further improve the trackers' performance especially under those challenging situations, such as low image resolution, severe illumination variation, serious motion blur, sudden pose change, heavy object occlusion, etc. We show some difficult cases in Fig. 1 .

Recently, some works [1, 2, 3] applied CNN networks in their algorithms by transfer learning. That is fine tune network models which have been trained on large scale image database-ImageNet [4]. Hong et al. [1] put an online Support Vector Machine (SVM) on top of hidden layers to compute target-specific saliency maps and predicted targets' position according to the saliency maps. In [2], Wang et al. conducted an in-depth study about the middle features produced by CNNs, and proposed a fully convolutional network combined general network and specic network for visual tracking. The network is a structured output CNN, and the output of it is a 50 50 probability map. In [3], Wang et al. treated each feature map channel as a base leaner and the CNN network was trained in the form of ensemble learning. The final results were also determined according to the output heat maps.

However, the works discussed above did not effectively exploit temporal context information which can play an important role in visual tracking. On the contrary, we proposed a novel generative model for visual tracking, it utilized the powerful CNN features directly to compute similarities between candidates and three kinds of target templates. We first modified the widely used VGG-M nets [5], then pretrained it in the same way as the authors described in [6], more details will be introduced in part 3. In the tracking process, we make use of the outputs of the top hidden convolution layer to compute similarities. In our algorithm, we consider three kinds of similarities which correspond to three kinds of important states as the way we have introduced in our other work [7]. The first kind of similarity is related to the most important targets' states in all

*Corresponding author
Email addresses: fengping@hust.edu.cn (Ping Feng), xuchunyan01@gmail.com (Chunyan Xu), zqzhao2000@foxmail.com (Zhiqiang Zhao), fang.liu@hust.edu.cn (Fang Liu), jj_guo@hust.edu.cn (Jingjuan Guo), yuanch@hust.edu.cn (Caihong Yuan), tjwang719@gmail.com (Tianjiang Wang), kuiduan@foxmail.com (Kui Duan)