Metric Learning Based Structural Appearance Model for Robust Visual Tracking

Yuwei Wu, Bo Ma, Min Yang, Jian Zhang, *Senior Member, IEEE,* and Yunde Jia, *Member, IEEE*

***Abstract*—Appearance modeling is a key issue for the success of a visual tracker. Sparse representation based appearance modeling has received an increasing amount of interest in recent years. However, most of existing work utilizes reconstruction errors to compute the observation likelihood under the generative framework, which may give poor performance, especially for significant appearance variations. In this paper, we advocate an approach to visual tracking that seeks an appropriate metric in the feature space of sparse codes and propose a metric learning based structural appearance model for more accurate matching of different appearances. This structural representation is acquired by performing multiscale max pooling on the weighted local sparse codes of image patches. An online multiple instance metric learning algorithm is proposed that learns a discriminative and adaptive metric, thereby better distinguishing the visual object of interest from the background. The multiple instance setting is able to alleviate the drift problem potentially caused by misaligned training examples. Tracking is then carried out within a Bayesian inference framework, in which the learned metric and the structure object representation are used to construct the observation model. Comprehensive experiments on challenging image sequences demonstrate qualitatively and quantitatively that the proposed algorithm outperforms the state-of-the-art methods.**

***Index Terms*—Appearance modeling, multiple instance metric learning, multiscale max pooling, object tracking, sparse coding.**

I. INTRODUCTION

APPEARANCE modeling is one of the most critical prerequisites for successful visual tracking. Designing an effective appearance model, however, is a challenging task due to appearance variations caused by background clutters, object deformation, partial occlusions, and illumination changes. In this paper, we address visual tracking by learning a suitable metric matrix in the feature space of local sparse codes to effectively capture appearance variations such that different appearances of an object will be close to each other and be well distinguished from the background simultaneously.

Sparse representation based appearance modeling has received considerable attention in the visual tracking community. The pioneer work introduced in [1] models the object appearance as a sparse linear combination of both object templates and trivial templates via ℓ_1 minimization. Most appearance models [1]–[5] based on sparse representation measure the similarity between the candidates and the models using reconstruction errors. The observation likelihood measured by the reconstruction error under the generative framework, however, suffers from following drawbacks.

1) The magnitude of reconstruction error depends largely on the accuracy of dictionary that is usually updated in an online manner to account for varying appearances. However, straightforward dictionary updating with newly obtained results is prone to drifting because of the accumulation of errors.
2) The measurement of reconstruction errors usually employs a predefined metric that lacks the ability to adapt to appearance changes.
3) Since most formulations of the observation likelihood do not take the background into account, they are less effective for tracking in cluttered environments due to the lack of discrimination.

To address the aforementioned shortcomings, employing local spare representation, we concentrate on seeking for an appropriate metric via online metric learning to find the region most similar to a given template, rather than locating the object with a minimal reconstruction error. We therefore present an effective metric learning based structural appearance model to alleviate the impacts of significant appearance variations, as shown in Fig. 1. Our method mainly includes two points.

1) A visual object of interest is characterized by performing multiscale pooling on the weighted local sparse coding. Sparse codes calculated from local image patches are susceptible to the slight variation (e.g., translation or rotation) and the noise. To alleviate this, the statistics of sparse codes using multiscale max pooling is computed to consider the spatial information of local descriptors and also to be invariant to translations.
2) An online metric learning algorithm is proposed to verify the most likely candidate. The updates of appearance models are sensitive to the training examples extracted

Manuscript received February 5, 2013; revised June 16, 2013; accepted August 23, 2013. Date of publication November 14, 2013; date of current version May 2, 2014. This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant 61375044, in part by the NSFC–Guangdong Joint Fund under Grant U1035004, in part by the Specialized Fund for Joint Building Program of Beijing Municipal Education Commission, and in part by the Specialized Research Fund for the Doctoral Program of Higher Education under Grant 20121101110035. This paper was recommended by Associate Editor R. Hamzaoui. (*Corresponding author: B. Ma.*)

Y. Wu, B. Ma, M. Yang, and Y. Jia are with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: bma000@126.com).

J. Zhang is with the Advanced Analytics Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2013.2291283