DOCTORAL THESIS

Multi-cue visual tracking: feature learning and fusion

Lan, Xiangyuan

Date of Award:
2016

Abstract

As an important and active research topic in computer vision community, visual tracking is a key component in many applications ranging from video surveillance and robotics to human computer interaction. Although it has been extensively studied in last two decades and significant progress has been made in recent years, it still remains challenging mainly due to numerous appearance variations caused by illumination, pose, occlusion, shape deformation and so on. Therefore, how to design an effective object appearance model which is able to handle different appearance changes has become one of the key problems in visual tracking. Since a single visual cue (feature) may not be sufficient to account for different appearance variations, multiple visual cues which describe different characteristics of the tracked object, e.g. color, texture, shape can be exploited jointly to achieve a more robust appearance model. In this thesis, we propose new appearance models based on multiple visual cues and address several research issues in feature learning and fusion for visual tracking.

Feature extraction and feature fusion are two key modules to construct the appearance model for the tracked target with multiple visual cues. Feature extraction aims to extract informative features for visual representation of the tracked target, and many kinds of hand-crafted feature descriptors which capture different types of visual information have been developed. However, since large appearance variations, e.g. occlusion, illumination may occur during tracking, the target samples may be contaminated/corrupted. As such, the extracted raw features may not be able to capture the intrinsic properties of the target appearance. Besides, without ex-

plicitly imposing the discriminability, the extracted features may potentially suffer background distraction problem. To extract uncontaminated discriminative features from multiple visual cues, this thesis proposes a novel robust joint discriminative feature learning framework which is capable of 1) simultaneously and optimally removing corrupted features and learning reliable classifiers, and 2) exploiting the consistent and feature-specific discriminative information of multiple feature. In this way, the features and classifiers learned from potentially corrupted tracking samples can be better utilized for target representation and foreground/background discrimination.

As shown by the Data Processing Inequality, information fusion in feature level contains more information than that in classifier level. In addition, not all visual cues/features are reliable, and thereby combining all the features may not achieve a better tracking performance. As such, it is more reasonable to dynamically select and fuse multiple visual cues for visual tracking. Based on aforementioned considerations, this thesis proposes a novel joint sparse representation model in which feature selection, fusion, and representation are performed optimally in a unified framework. By taking advantages of sparse representation, unreliable features are detected and removed while reliable features are fused on feature level for target representation. In order to capture the non-linear similarity of features, the model is further extended to perform feature fusion in kernel space. Experimental results demonstrate the effectiveness of the proposed model.

Since different visual cues extracted from the same object should share some commonalities in their representations and each feature should also have some diversities to reflect its complementarity in appearance modeling, another important problem in feature fusion is how to learn the commonality and diversity in the fused representations of multiple visual cues to enhance the tracking accuracy. Different from existing multi-cue sparse trackers which only consider the commonalities among the sparsity patterns of multiple visual cues, this thesis proposes a novel multiple sparse representation model for multi-cue visual tracking which jointly exploits

the underlying commonalities and diversities of different visual cues by decomposing multiple sparsity patterns. Moreover, this thesis introduces a novel online multiple metric learning to efficiently and adaptively incorporate the appearance proximity constraint, which ensures that the learned commonalities of multiple visual cues are more representative. Experimental results on tracking benchmark videos and other challenging videos show that the proposed tracker achieves better performance than the existing sparsity-based trackers and other state-of-the-art trackers.

In short, the major contributions of this thesis are summarized as follows.

- A robust feature learning model is proposed to learn uncontaminated and discriminative features from multiple visual cues for feature extraction in visual tracking.

- A robust feature fusion model based on joint sparse representation is proposed to dynamically select and fuse multiple visual cues on feature level for visual tracking.

- A multiple sparse representation framework is proposed to explicitly model the commonality and diversity in the representation of multiple visual cues, which achieves more accurate appearance modeling with multiple features in visual tracking.

Keywords: visual tracking, feature learning, feature fusion

Table of Contents