Video Tracking Using Learned Hierarchical Features

mingzailao

2016-9-11

Outline

ASLSA(adaptive structural local sparse appearance model) [1]

Tracking System Overview

Briefly Introduction of the Tracking System

Suppose we have an observation set of target $x_{1:t} = \{x_1, \cdots, x_t\}$, a corresponding feature representation set $z_{1:t} = \{z_1, \cdots, z_t\}$, the target state y_t can be calculated by:

$$y_t = arg \max_{y_t^i} p(y_t^i | z_{1:t}) \tag{1}$$

where y_t^i denotes the i^{th} sample in the t^{th} frame.

Tracking System Overview

Briefly Introduction of the Tracking System

The posterior probability $p(y_t|z_{1:t})$ can be inferred by the Bayes theorem as follows:

$$p(y_t|z_{1:t}) \propto p(z_t|y_t) \int p(y_t|y_{t-1}) p(y_{t-1}|z_{1:t-1}) \tag{2}$$

where $z_{1:t}$ denotes the feature representation, $p(y_t|y_{t-1})$ denotes the motion model and $p(z_t|y_t)$ denotes the appearance model.

Tracking System Overview

Briefly Introduction of the Tracking System

The representations $z_{1:t}$ can simply use raw pixel values. [1] In there , we use the learned hierarchical features from raw pixels for tracking.

Learning Features for Video Tracking

Offline Learning

- Adopt the approach proposed in [2] to learn features From a auxiliary dataset.
- We further use a domain adaptation method to adapt pre-learned features according to specific target objects.

Learning Features for Video Tracking

Algorithm

- Input: the previous tracking state y_{t-1}, the existing feature learning parameter $\hat{\Theta}$ and the exemplar library.
- Apply the affine transformation on y_{t-1} to obtain a number of tracking states y_t^i and the corresponding candidate image patches x_t^i.
- Extract feature representations z_t^i from the candidate image patches x_t^i under the existing feature learning parameter $\hat{\Theta}$.

Learning Features for video Tracking

Algorithm

- Calculate the posterior probability $p(y_t|z_{1:t})$ according to Equation (2).
- Predict the tracking state by $y_t = arg\max_{y_t^i} p(y_t^i|z_{1:t})$.
- Update the feature learning parameter and the exemplar library every M frames.
- Output: the predicted tracking state y_t, the up-to-date feature learning parameter Θ and the up-to-date exemplar library.

Pre-Learning Generic Features from Auxiliary Videos

Deep Learning model

Network Structure [2]

Reference

Xu Jia, Huchuan Lu, and Ming-Hsuan Yang.
Visual tracking via adaptive structural local sparse appearance model.
In *Computer vision and pattern recognition (CVPR), 2012 IEEE Conference on*, pages 1822–1829. IEEE, 2012.

Will Zou, Shenghuo Zhu, Kai Yu, and Andrew Y Ng.
Deep learning of invariant features via simulated fixations in video.
In *Advances in neural information processing systems*, pages 3212–3220, 2012.