Computer Vision and Image Understanding 000 (2016) 1–11

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Online adaptive hidden Markov model for multi-tracker fusion

Tomas Vojir[a,*], Jiri Matas[a], Jana Noskova[b]

[a] The Center for Machine Perception, FEE CTU in Prague, Karlovo namesti 13, 121 35 Prague 2, Czech Republic
[b] Faculty of Civil Engineering, CTU in Prague, Thakurova 7/2077, 166 29 Prague 6, Czech Republic

ARTICLE INFO

Article history:
Received 27 July 2015
Revised 21 January 2016
Accepted 15 May 2016
Available online xxx

Keywords:
Visual tracking
On-line learning
Hidden markov model
Object detection

ABSTRACT

In this paper, we propose a novel method for visual object tracking called HMMTxD. The method fuses observations from complementary out-of-the box trackers and a detector by utilizing a hidden Markov model whose latent states correspond to a binary vector expressing the failure of individual trackers. The Markov model is trained in an unsupervised way, relying on an online learned detector to provide a source of tracker-independent information for a modified Baum- Welch algorithm that updates the model w.r.t. the partially annotated data. We show the effectiveness of the proposed method on combination of two and three tracking algorithms. The performance of HMMTxD is evaluated on two standard benchmarks (CVPR2013 and VOT) and on a rich collection of 77 publicly available sequences. The HMMTxD outperforms the state-of-the-art, often significantly, on all data-sets in almost all criteria.

1. Introduction

In the last thirty years, a large number of diverse visual tracking methods has been proposed (Smeulder et al., 2013; Yilmaz et al., 2006). The methods differ in the formulation of the problem, assumptions made about the observed motion, in optimization techniques, the features used, in the processing speed, and in the application domain. Some methods focus on specific challenges like tracking of articulated or deformable objects (Cehovin et al., 2013; Godec et al., 2011; Kwon and Lee, 2009), occlusion handling (Grabner et al., 2010), abrupt motion (Zhou and Lu, 2010) or long-term tracking (Kalal et al., 2012; Pernici and Bimbo, 2013).

Three observations motivate the presented research. First, most trackers perform poorly if run outside the scenario they were designed for. Second, some trackers make different and complementary assumptions and their failures are not highly correlated (called complementary trackers in the paper). And finally, even fairly complex well performing trackers run at frame rate or faster on standard hardware, opening the possibility for multiple trackers to run concurrently and yet in or near real-time.

We propose a novel methodology that exploits a hidden Markov model (HMM) for fusion of non-uniform observables and pose prediction of multiple complementary trackers using an on-line learned high-precision detector. The non-uniform observables, in this sense, means that each tracker can produce its own type of "confidence estimate" which may not be directly comparable between each other.

The HMM, trained in an unsupervised manner, estimates the state of the trackers – failed, operates correctly – and outputs the pose of the tracked object taking into account the past performance and observations of the trackers and the detector. The HMM treats the detector output as correct if it is not in contradiction with its current most probable state in which the majority of trackers are correct. This limits the cases where the HMM would be wrongly updated by a false detection. For the potentially many frames where reliable detector output is not available, it combines the trackers. The detector is trained on the first image and interacts with the learning of the HMM by partially annotating the sequence of HMM states in the time of verified detections. The recall of the detector is not critical but it affects the learning rate of the HMM and the long-term properties of the HMMTxD method, i.e., its ability to re-initialize trackers after occlusions or object disappearance.

Related work. The most closely related approaches include Santner et al. (2010), where three tracking methods with different rates of appearance adaptation are combined to prevent drift due to incorrect model updates. The approach uses simple, hard-coded rules for tracker selection. Kalal et al. (2012) combine a tracking-by-detection method with a short-term tracker that generates so called P-N events to learn new object appearance. The output is defined either by the detector or the tracker based on visual similarity to the learned object model. Both these methods employ pre-defined rules to make decisions about

* Corresponding author.
E-mail addresses: vojirtom@cmp.felk.cvut.cz (T. Vojir), matas@cmp.felk.cvut.cz (J. Matas), noskova@fsv.cvut.cz (J. Noskova).
URL: http://cmp.felk.cvut.cz/~vojirtom/ (T. Vojir), http://cmp.felk.cvut.cz/~matas/ (J. Matas)

http://dx.doi.org/10.1016/j.cviu.2016.05.007
1077-3142/

object pose and use one type of measurement, a certain form of similarity between the object and the estimated location. In contrary, HMMTxD learns continuously and causally the performance statistics of individual parts of the systems and fuses multiple "confidence" measurements in the form of probability densities of observables in the HMM. Zhang et al. (2014) use a pool of multiple classifiers learned from different time spans and choose the one that maximize an entropy-based cost function. This method addresses the problem of model drifting due to wrong model updates, but the failure modes inherent to the classifier itself remains the same. This is unlike the proposed method which allows to combine diverse tracking methods with different inherent failure modes and with different learning strategies to balance their weaknesses.

Similarly to the proposed method, Wang and yan Yeung (2014) and Bailer et al. (2014) fuse different out-of-the box tracking methods. Bailer et al. combine offline the *outputs* of multiple tracking algorithms. There is no interaction between trackers, which for instance implies that the method avoids failure only if one method correctly tracks the whole sequence. Wang et al., use a factorial hidden Markov model and a Bayesian approach. The state space of their factorial HMM is the set of potential object positions, therefore it is very large. The model contains a probability description of the object motion based on a particle filter. Trackers interact by reinitializing those with low reliability to the pose of the most confident one. The Yuan et al. (2015) use HMM in the same setup, but rather than merging multiple tracking method, they focus on modeling the temporal change of the target appearance in the HMM framework by introducing a observational dependencies. In contrast, the HMMTxD method is online with tracker interaction via a high precision object detector that supervises tracker re-initializations which happen on the fly. The appearance modeling is performed inside of each tracker and the HMMTxD capture the relation of the confidence provided by tracker and its performance, validated by the object detector, by the observable distributions. Moreover, the HMMTxD confidence estimation is motion-model free and this prevents biases towards support of trackers with a particular motion model.

Yoon et al. (2012) combines multiple trackers in a particle filter framework. This approach models observables and transition behavior of individual trackers, but the trackers are self-adapting which makes it prone to wrong model updates. The adaptation of HMMTxD model is supervised by a detector method set to a specific mode of operation – near 100% precision – alleviating the incorrect update problem.

The contributions of the paper are: a novel method for fusion of multiple trackers based on HMMs using non-uniform observables, a simple, and so far unused, unsupervised method for HMMs training in the context of tracking, tunable feature-based detector with very low false positive rate, and the creation of a tracking system that shows state-of-the-art performance.

2. Fusing multiple trackers

HMMTxD uses a hidden Markov model (HMM) to integrate pose and observational confidence of different trackers and a detector, and updates its own confidence estimates that in turn define the pose that it outputs. In the HMM, each tracker is modeled as working correctly (1) or incorrectly (0). The HMM poses no constraints on the definition of tracker correctness, we adopted target overlap above a threshold. Having at our disposal $\mathbf{n}$ trackers, the set of all possible states is $\{s_1, s_2, \ldots, s_N\} = \{0, 1\}^{\mathbf{n}}, N = 2^{\mathbf{n}}$ and the initial state $s_1 = (1, 1, \ldots, 1)$. Note that the trackers are not assumed to be independent, because an independence of tracker correctness is not a realistic assumption. For example, if the tracking problem is relatively easy, all trackers tend to be correct and in the case of occlusion all tend to be incorrect (see the analysis in Kristan et al., 2015). The number of states $2^{\mathbf{n}}$ grows exponentially with the number of trackers. However, we do not consider this a significant issue – due to "real-time" requirements of tracking, the need to combine more than a small number of trackers, say $\mathbf{n} = 4$, is unlikely.

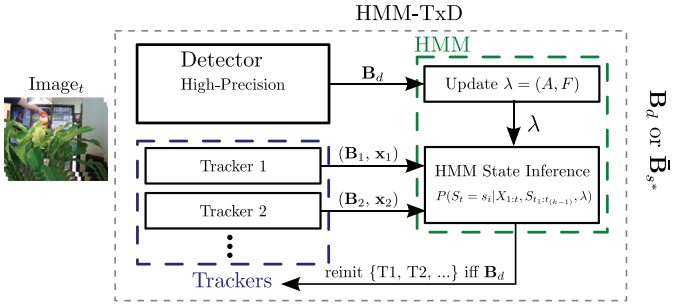

Fig. 1. The structure of the HMMTxD. For each frame, the detector and trackers are run. Each tracker outputs a new object pose and observables ($\mathbf{B}_i$, $\mathbf{x}_i$) and the detector outputs either the verified object pose $\mathbf{B}_d$ or nothing. If detector fires, HMM is updated and trackers are re-initialized and the final output is the $\mathbf{B}_d$, otherwise, HMM estimate the most probable state s^* and outputs an average bounding box $\bar{\mathbf{B}}_{s^*}$ of trackers that are correct in the estimated state s^*.

The HMMTxD method overview is illustrated in Fig. 1. Each tracker provides an estimate of the object pose ($\mathbf{B}_i$) and a vector of observables ($\mathbf{x}_i$), which may contain a similarity measure to some model (such as normalized cross-correlation to the initial image patch, distance of template and current histograms at given position, etc.) or any other estimates of the tracker performance. The $\mathbf{x}_i, i = \{1, 2, \ldots, \mathbf{n}\}$ serve as observables to relate the tracker current confidence to the HMM. Each individual observable depends only on one particular tracker and its correctness, hence, they are assumed to be conditionally independent conditioned on the state of the HMM (which encodes the tracker correctness).

In general, there are no constraints on observable values, however, in the proposed HMM the observable values are required to be normalized to the (0, 1) interval. The observables are modeled as beta-distributed random variables (Eq. 1) and its parameters are estimated online. The beta distribution was chosen for its versatility, where practically any kind of unimodal random variable on (0, 1) can be modeled by the beta distribution, i.e., for any choice of any lower and upper quantiles, a beta distribution exists satisfying the given quantile constraint (Gupta and Nadarajah, 2004).

Learning the parameters of the beta distributions online is crucial for the adaptability to particular tracking scenes, where the observable values from a different trackers may be biased due to scene properties, or to adapt to a different types of observables of trackers and their correlations to the "real" tracker performance. For example, taking correlation with the initial target patch as an observable for one tracker and color histogram distance to a initial target for a second tracker, the correlation between their values and the performance of the tracker may differ depending on object rigidity and color distribution of object and background.

The HMM is parameterized by the pair $\lambda = (A, F)$, where A are the probabilities of state transition and F are the beta distributions of observables with shape parameters $p, q > 0$ and density defined for $x \in (0, 1)$

$$f(x|p, q) = \frac{x^{p-1}(1-x)^{q-1}}{\int_0^1 u^{p-1}(1-u)^{q-1}du}. \quad (1)$$

Since the goal is real-time tracking without any specific preprocessing, learning of HMM parameters has to be done online. Towards this goal, the object detector, which is set to operating mode with low false positive rate, is utilized to partially annotate the sequence of hidden states. In contrast to classical HMM, where only a

Please cite this article as: T. Vojir et al., Online adaptive hidden Markov model for multi-tracker fusion, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.05.007