IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

An Efficient Difficult Appearance of Object Tracking Based on Spatial-Temporal Consistency

Dr. R. Muthu Kumar[1], M.Sivakumar[2]

Associate Professor, Shree Venkateshwara HI-Tech Engineering College, Gobi, India[1]

PG Student (ME-AE), Shree Venkateshwara HI-Tech Engineering College, Gobi, India[2]

Abstract: Analysing image sequences to detect and determine temporal events is often known as video analysis. The "video analysis is used in a wide range of domains including entertainment, health care, automotive, transport, home automation, safety and security. Object detection is the first and most important step of moving object tracking. These techniques can be divided into several categories. Moving objects extraction from background is an important task. Results of object extraction depend upon the variation of local or global light intensities, objects shadow, background and foreground regular or irregular movement. Present a robust tracking method by exploiting a fragment-based appearance model with consideration of both temporal continuity and discontinuity information. From the perspective of probability theory, the proposed tracking algorithm can be viewed as a two-step optimization problem. In the first step, by adopting the estimated occlusion state as a prior, the optimal state of the tracked object can be obtained by solving an optimization problem, where the objective function is designed based on the classification score, occlusion prior, and temporal continuity information. In the second step propose a discriminative occlusion model, which exploits both foreground and background information to detect the possible difficult appearance, and also models the consistency of occlusion labels among different frames. In addition, a simple yet effective training strategy is introduced during the model training (and updating) process, with which the effects of spatial-temporal consistency are properly weighted. The propose tracker is evaluated by using the recent benchmark data set, on which the results demonstrate that our tracker performs favourably against other state-of-the-art tracking algorithms.

Keywords: Spatial Reasoning, Visual Surveillance, Temporal Reasoning.

I. INTRODUCTION

The fundamental problems in computer vision, object tracking plays a critical role in numerous lines of research such as motion analysis, image compression, and activity recognition. While much progress has been made in the past decades, developing a robust online tracker is still a challenging problem due to difficulties to account for appearance change of a target object, which includes intrinsic (e.g., pose variation and shape deformation) and extrinsic factors (e.g., varying illumination, camera motion, and occlusions).

Visual tracking, as a fundamental problem in computer vision, has found wide applications. Although much progress has been made in the past decade, tremendous challenges still exist in designing a robust tracker that can well handle significant appearance changes, pose variations, severe occlusions, and background clutters. Existing appearance-based tracking methods adopt either generative or discriminative models to separate the foreground from background and distinct co-occurring objects. One major drawback is that they rely on low-level hand-crafted features which are incapable to capture semantic information of targets, not robust to significant appearance changes, and only have limited discriminative power. A tracking method typically consists of three components: an appearance (observation) model which evaluates the likelihood of an observed image patch (associated to a state) belonging to the object class; a dynamic model (or motion model), which aims to describe the states of an object over time.

Visual tracking is to identify the position of a pre-specified object continuously in a given image sequence, which is very attractive in computer vision for its wide applications in numerous domains. Nowadays, it still remains a challenging task to develop a sophisticated algorithm which performs object tracking robustly and accurately due to many difficulties. These difficulties are mainly attributed to appearance changes, severe occlusion, illumination variations, and background.

The tracking method typically consists of three components: an appearance (observation) model which evaluates the likelihood of an observed image patch (associated to a state) belonging to the object class; a dynamic model (or motion model), which aims to describe the states of an object overtime (e.g., Kalman filter and particle filter); and a search strategy for finding the likely states in the current frame (e.g., mean shift and sliding window). In this paper, we propose a robust appearance model that considers the effects of occlusion and motion blur. Hence, we only discuss key issues related to appearance models rather than present a detailed review of all components.

DOI10.17148/IJARCCE.2017.63153

In order to develop effective appearance models for robust object tracking, several critical factors need to be considered. The first one is concerned with how objects are represented.

Any representation scheme can be categorized based on adopted features (e.g., intensity, colour, texture, Haar- like feature, super-pixel based feature, and sparse coding), and description models (e.g., holistic histogram, part-based histogram, and subspace representation). Instead of treating the target object as a collection of low-level features, subspace representation methods provide a compact notion of the "thing" being tracked, which facilitates other vision tasks (e.g., object recognition).

II. RELATED WORK

Cognitive science and cognitive neuroscience aim at understanding and clarifying human cognition and, in the last decades, the Signal Processing community has experienced fruitful fertilization from such disciplines. In particular, of primary interest is the human capability of adapting to the new situations. This feature can be very valuable especially in non-stationary stochastic environments and can be seen as the result of the actuation of the so called cognitive cycle. Every step (Sensing, Analysis, Decision and Action) is linked to a learning phase. These concepts have been lately applied to the computer vision research field, aiming to design more robust, resilient and adaptable computer vision systems, by mimicking the human capabilities.

A. Object Tracking with Incremental Subspace Learning

Object tracking via online subspace learning has attracted much attention in recent years. The incremental visual tracking (IVT) method introduces an online update approach for efficiently learning and updating a low dimensional PCA subspace representation of the target object. Several experimental results demonstrate that PCA subspace representation with online update is effective in dealing with appearance change caused by in-plane rotation, scale, illumination variation and pose change. However, it has also been shown that the PCA subspace based representation scheme is sensitive to partial occlusion, which can be where y denotes an observation vector, z indicates the corresponding coding or coefficient vector, U represents a matrix of column basis vectors, and e is the error term.

B. Object learning-based method

Temporal continuity (usually called motion model) aims to depict the state transition of the tracked object over time, which is a very important characteristic in visual tracking. In recent "particle filter"-based trackers, they usually assume that the motion of the tracked object between two consecutive frames follows a Gaussian distribution. This assumption is too rough as it does not take motion estimation into account. Thus, it limits the tracking performance especially when some large or complex motions occur. Although the motion estimation is well studied in video segmentation methods, it has not been paid much attention to in the tracking field. Introduce temporal continuity into a shape model, and estimate the motion of the shapes between two consecutive frames by computing an affine transform. The temporal consistent segmentations of moving objects are achieved by clustering the point trajectories which have different survival time. For visual tracking cast motion estimation as a linear assignment problem based on a set of short exploit the optical flow estimation across frames, based on which a thresholder motion mode lis designed to describe the motion of the tracked object. This motion model penalizes motions that violate the optical flow estimation, and therefore makes the tracker more effective in handling many complex challenges (such as out of plane rotation, deformation and so on).

However, this threshold motion model is very dependent on the accurate optical flow estimation. If the optical flow is not accurate, the thresholding operator may lead to completely wrong motion estimation. In this work, we also introduce the optical estimation into our temporal continuity model, and determine its contribution by using a learning-based method (i.e., a linear SVM classifier) rather than adopting a simple thresholding operator.

C. Object learning-based on Machine learning

Object detectors are traditionally trained assuming that all training examples are labeled. Such an assumption is too strong in our case since we wish to train a detector from a single labeled example and a video stream. This problem can be formulated as a semi-supervised learning that exploits both labeled and unlabeled data. These methods typically assume independent and identically distributed data with certain properties, such as that the unlabeled examples form "natural" clusters in the feature space. A number of algorithms relying on similar assumptions have been proposed in the past including EM, Self-learning and Co-training. Expectation-Maximization (EM) is a generic method for finding estimates of model parameters given unlabeled data. EM is an iterative process, which in case of binary classification alternates over estimation of soft-labels of unlabeled data and training a classifier. EM was successfully applied to document classification and learning of object categories].In the semi-supervised learning terminology, EM algorithm relies on the" low density separation" assumption, which means that the classes are well separated. EM is sometimes interpreted as a "soft" version of self-learning.

D. Most related approaches

Many approaches combine tracking, learning and detection in some sense. An offline trained detector is used to validate the trajectory output by a tracker and if the trajectory is not validated, an exhaustive image search is performed to find the target. Other approaches integrate the detector with in a particle filtering framework. Such techniques have been applied to tracking of faces in low

DOI10.17148/IJARCCE.2017.63153

frame-rate video, multiple hockey players, or pedestrians. In contrast to our method, these methods rely on an offline trained detector that does not change its properties during run-time. Adaptive discriminative trackers also have the capability to track, learn and detect. These methods realize tracking by an online learned detector that discriminates the target from its background. In other words, a single process represents both tracking and detection. This is in contrast to our approach where tracking and detection are independent processes that exchange information using learning. By keeping the tracking and detection separated our approach does not have to compromise neither on tracking nor detection capabilities of its components.

III.PROPOSED APPROACH

The proposed a novel fragment-based tracking algorithm by using a two-step optimization framework, which takes both temporal discontinuity and spatial-temporal continuity into consideration simultaneously. First, with the estimated occlusion labels of the target's different fragments, the spatial-temporal structure of the tracked object is obtained by solving an optimization problem that can be handled effectively by the dynamic programming method in a recursive manner. The second stage is to estimate occlusion labels if the optimal states (i.e., spatial-temporal structure) are obtained. In this stage, a simple yet effective L2-distance matching method is proposed, which makes the occlusion model rapidly adapt to scene changes. For model training (and updating), It extract positive and negative samples based on the tracking result, and update our tracker every five frames.

The proposed approach following steps is followed.

- A unified tracking framework is proposing, in which the robust appearance description, spatial consistency, temporal continuity and temporal discontinuity are fully considered. First, consider the spatial and temporal configurations by introducing a spatial-temporal tree. The relations between spatial and temporal fragments are well considered and effectively trained. Second, the temporal continuity model recursively estimates the motion consistency among different frames, in which a learning-based transfer score function is designed to consider the guidelines of the optical flow estimation. In addition, the temporal discontinuity model enables tracker to robustly handle occlusions by inferring the occlusion states explicitly. With the estimated occlusion state, the tracker can also ignore the inaccurate motion estimation and therefore avoid error accumulation.
- A two-step optimization manner is developing. In each step, the recursive energy function is derived to model different components. In propose tracking framework, we firstly adopt a joint distribution to depict both locations and occlusion states. Then, we exploit a two-step optimization strategy to infer the optimal locations and best occlusion states separately rather than solve them jointly. For one thing, the model parameters (i.e., classification coefficients) in the joint distribution are updated moderately to avoid model degradation and tracking drift
- Promising tracking performance is achieved in theCVPR2013 bench mark. The propose system conduct extensive experiments to compare our tracker with numerous state-of-the-art methods on the recent benchmark dataset. It can be seen from experimental results that the proposed tracker performs favourably against the compared state-of-the-art trackers.

A. Fragment-Based Appearance Model

First, introduce a fragment-based appearance model and define a score function to depict. It $S_A(B_t) = \mathrm{lpg}\, P(Y_t|B_t o_{1:t-1})$ define the score function as,

$$S_A(B_t) = \sum_{j=0}^{M} f_A\left(b_t^j\right) + \lambda_1 \sum_{j=1}^{M} g_A(b_t^0, b_t^j) \quad (1)$$

Where feature scores for different fragments$f_A\left(b_t^j\right)$ $j \in \{0, \ldots . M\}$.

The second term in equation (1) denotes the spatial score, which models spatial relationships between different fragments (except fragment 0) and the holistic template (fragment 0),in which the basic function gA (·) is defined as,

$$g_A\left(b_t^0, b_t^j\right) = \langle V_t^j, \phi_s\ (b_t^0, b_t^j)\rangle \quad (2)$$

Where$\phi_s\ (b_t^0, b_t^j)$is the spatial relation vector between fragment 0 and fragment j that can be calculated by

$$\phi_s\left(b_t^0, b_t^j\right) = [dx, dx^2, dy\, dy^2](3)$$

And V_t^jdenotes the corresponding classification coefficient vector. The vectors V_t^j $j \in (1, \ldots M)$, depict the spatial configurations between fragment 0 and fragment j, which are online updated every frame (i.e., the relative positions of fragments are not fixed in this work).

B. Temporal Consistency Model

In many recent trackers, motions between consecutive frames are usually assumed as a Gaussian random walk process. This assumption limits the tracking performance for two main reasons: (1) it is not very accurate to model the object motion as a Gaussian distribution (see Appendix B for more discussions); (2) information from previous frames is lost as only two consecutive frames are considered. In this paper, we address these problems by using a recursive temporal consistency model.

Recall the second term in equation (3), where the probability$p(B_{1:t}|Y_{1:t-1,}o_{1:t-1})$ can be easily decomposed as

$$p(B_{1:t}|Y_{1:t-1}) \propto p(B_{1:t-1}|Y_{1:t-2} o_{1:t-1}) \times p(B_t|B_{t-1} o_{1:t-1}) \times p(Y_{t-1}|B_{t-1} o_{1:t-1} \quad (4)$$

The proposed work, we determine the joint distribution of bounding box set $B_{1:t-1}$ and the observation likelihood of $B_{t-1}$with occlusion labels in previous frames (i.e., o1:t−2). By taking the logarithm operator at both sides of equation (4), it can obtain the following equation,

$$S_T(B_{1:t}) = S_T^1(B_{t-1}) + S_T^2(B_t, B_{t-1}) + S_T(B_{1:t-1}) + C_t \quad (5)$$

This term is exploited to measure the compatibility between the optical flow and the estimated object motion, in which pjt stands for the estimated displacement vector between frame t and frame t − 1 via the optical flow method.

C. Model Inference

After SA (·) and ST (·) are determined, we can obtain.$\widehat{B_t}$To be specific, we adopt the dynamic programming algorithm to perform this optimization process. The dynamic programming algorithm defined on a tree, which aims to solve the following optimization problem".

$$L^* = arg\ min_L \left(\sum_{p=1}^{N} m_p(l_p) + \sum_{(v_p, v_q)} d_{pq}(l_p, l_q) \right) \quad (6)$$

Where $m_p(\cdot)$ and$d_{pq}(l_p, l_q)$ denote the unary and binary terms,$L = (l_1 \dots . l_n)$ denotes a configuration of the tree structured model.

D. Scale Estimation

In exiting tracking process just tries to estimate the optimal target position, and does not take target scale into consideration. The proposed system further refines the tracking results by estimating the target scale. Specially, it first estimates the displacement (i.e., centre position) of the target via the above method. Then, refine the scale of it by sampling candidate scales on the basis of the estimated centre and the previous scale.

The affine parameter$A = [x_t, y_t, s_t, \theta_t, \alpha_t, \phi_t]$ to denote the estimated bounding box , where $x_t, y_t, s_t, \theta_t, \alpha_t, \phi_t$ denote the x, y translation, scale, rotation angle, aspect ratio, and skew direction. We generate Ns vectors$P_n = [0,0, \delta_n, 0,0]|_{n=1}^{N_s}$, and then states of Ns candidate boxes are obtained as

$$A_n = A + P_n, \quad n = 1, \dots . N_s \quad (7)$$

To obtain the best state from the candidates via

$$\hat{A} = arg\ max_{A_n} \langle u_t, \Phi(A_n) \rangle \quad (8)$$

Where u_t denotes the SVM coefficients, $\Phi(A_n)$denotes feature vector extracted in the bounding box specified by An. Once the scale for one frame is determined, it fixes it for the following procedures.

E. Occlusion Model

In this work, the occlusion model is developed in a discriminative manner rather than a generative manner, i.e., model the posterior probability $p(o_{1:t}|Y_{1:t}B_{1:t}$directly rather than model the likelihood and prior separately. To be specific, here introduce an occlusion score function.

$$S_o(o_{1:t}) = \log p(o_{1:t}|Y_{1:t}B_{1:t}) \quad (9)$$

to model the occlusion labels of different fragments. By assuming that occlusion labels of different fragments are mutually independent3, we obtain the following equation,

$$S_o(o_{1:t}) = \sum_{j=0}^{M} e_o(o_{1:t}^j) \quad (10)$$

$e_o(o_{1:t}^j)$Denotes the occlusion score of the j -th fragment in frame t.

F. SVM Classifier

Update SVM Classifier: This proposed system updates the concatenated SVM coefficients $W_{t+1} = [u_{t+1}^0; \dots . u_{t+1}^M; c_{t+1}^0; \dots .\ c_{t+1}^M; v_{t+1}^1; \dots . v_{t+1}^M]$every five frames. As discussed in [15], computing W_{t+1}simultaneously may lead to the over-fitting problem.

Thus, in this work it divides the training (and updating) process into two kinds of sub-problems. First, it concatenates the feature vector of fragment0 with the relation vectors $\phi_s(.)\phi_t(.)$ and compute the classifier coefficients, i.e., $[u_{t+1}^0; \dots . v_{t+1}^1; v_{t+1}^2; \dots .\ v_{t+1}^M; c_{t+1}^0]$ .Second, we compute the classifier coefficients for the feature vector that is generated by concatenating the feature vector of fragment j with its connected temporal relation vector $\phi_t(.)$,i.e$[u_{t+1}^j, c_{t+1}^j]$$j \in \{1,,M\}$. We extract samples whose occlusion labels are 0, and combine them with old support vectors to form the feature pool n. Then the training (and updating) process is formulated as.

$$\arg\min_{\beta, \xi_n} \frac{1}{2} \beta^T \beta + C \sum \xi_n \quad (11)$$

Where β is the classifier coefficient of the concatenated feature, and K is the set of coefficients corresponding to the feature. In each sub-problem, it directly exploits the tracked configurations in previous frames as positive samples. For generating the negative samples, we generate configurations by moving the position of the holistic fragment (or fragment j in the second training process) from its tracked position and remain the spatial/temporal relationships among the fragments. Then the negative samples are extracted based on these newly generated configurations.

IV. EXPERIMENTAL RESULTS

The proposed tracking algorithm is implemented in MATLAB on a PC machine with Intel i7-6770 CPU(3.4 GHz) and 32 GB memory. The tracked object is divided into three fragments vertically or horizontally according to aspect ratio of the target. Specially, according to the annotation in the first frame of each sequence, we divide the target into three fragments horizontally when the width is1.7 times larger than the height, and vertically otherwise. Basically, we adopt the HOG (Histogram of Gradient) features to describe each divided fragment and the holistic template, which are normalized to unit vectors. For the tracker based on HOG features, it performs the target search within a radius of 30 pixels, which is the same as the radius in the negative samples extraction process. The parameters λ1, λ2 and α a irrespectively set to 0.005, 0.001 and 0.2. The number of frames we trace back in the inference process for both bounding box and occlusion label is set to 5. In addition to the HOG features, it also implements a version of our tracker by exploiting the color features.

Table 4.1: Comparison of Success Rate with object overlap

Techniques	**Overlap Threshold**					
	0.1	**0.2**	**0.4**	**0.6**	**0.8**	**1**
Cognitive Control	0.73	0.71	0.67	0.49	0.23	0.19
Spatial-Temporal Consistency	0.94	0.86	0.77	0.65	0.59	0.23

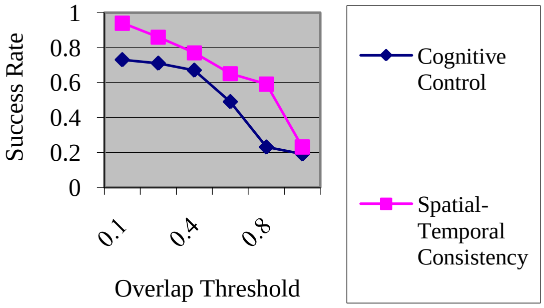

Fig 4.1 Comparison of different threshold setting with success rate analysis

Table 4.2 Comparison of precision with object location Error Threshold

Techniques	Location Error Threshold					
	10	20	30	40	50	60
Cognitive Control	0.25	0.48	0.54	0.61	0.78	0.81
Spatial-Temporal Consistency	0.27	0.55	0.69	0.74	0.82	0.93

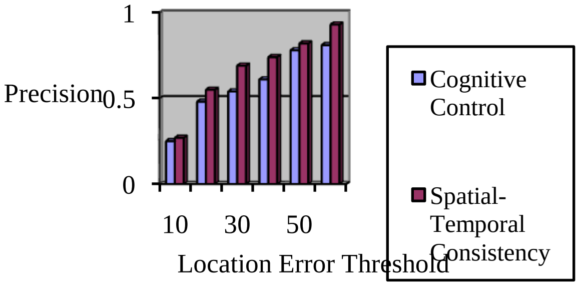

Fig 4.2 Comparison of location error with precision value

V. CONCLUSION

The proposed novel tracking algorithm by estimating the position of the tracked object and occlusion state in an iterative manner. With the computed occlusion, prior, a recursive inference is performed in the spatial-temporal structure, by which the positions of the tracked object (and fragments) of several frames are simultaneously optimized. For conducting occlusion estimation, a discriminative occlusion model is proposed, which directly compares the target with the positive and negative samples with an L2-norm distance. The temporal consistency of occlusion states among frames is also taken into consideration for optimizing. In addition, a simple yet effective training (and updating) strategy is also introduced to ensure the model coefficients are properly learned.

REFERENCES

[1] K. Granström, C. Lundquist, and U. Orguner, “A Gaussian mixture PHD filter for extended target tracking,” in Proc. 13th Conf. Inf. Fusion (FUSION), Jul. 2010, pp. 1–8.

[2] H. Firouzi and H. Najjaran, “Real-time monocular vision-based object tracking with object distance and motion estimation,” in Proc. IEEE/ASME Int. Conf. Adv. Intell. Mechatron. (AIM), Jul. 2010, pp. 987–992.

[3] Z. Kalal, K. Mikolajczyk, and J. Matas, “Tracking-learning-detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[4] J. Gao, H. Ling, W. Hu, and J. Xing, “Transfer learning based visual tracking with Gaussian processes regression,” in Computer Vision (Lecture Notes in Computer Science), vol. 8691, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Zurich, Switzerland: Springer, 2014, pp. 188–203.

[5] D. Wang, H. Lu, Z. Xiao, and M.-H. Yang, “Inverse sparse tracker with a locally weighted distance metric,” IEEE Trans. Image Process., vol. 24, no. 9, pp. 2646–2657, Sep. 2015.

[6] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High-speed tracking with kernelized correlation filters,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 583–596, Mar. 2015.

[7] D. Wang, H. Lu, and C. Bo, “Visual tracking via weighted local cosine similarity,” IEEE Trans. Cybern., vol. 45, no. 9, pp. 1838–1850, Sep. 2015.

[8] A. Mazzù, S. Chiappino, L. Marcenaro, and C. S. Regazzoni, “A switching fusion filter for dim point target tracking in infra-red video sequences,” in Proc. 17th Int. Conf. Inf. Fusion (FUSION), Jul. 2014, pp. 1–6.

[9] D. Wang, H. Lu, and M. H. Yang, “Robust visual tracking via least soft threshold squares,” IEEE Trans. Circuits Syst. Video Technol., vol. PP, no. 99, p. 1, 2015, doi: 10.1109/TCSVT.2015.2462012.

[10] L. Zhang, D. Zhang, Y. Su, and F. Long, “Adaptive kernel-bandwidth object tracking based on mean-shift algorithm,” in Proc. 4th Int. Conf. Intell. Control Inf. Process. (ICICIP), Jun. 2013, pp. 413–416.

BIOGRAPHIES

R. MuthuKumar received his B.E Degree in Electricaland Electronics Engg. from Bharathiyar University, Coimbatore, TamilNadu in the year 2002, M.E., in Power Systems Engg. From GCT, Coimbatore TamilNadu in the year 2006 and pursuing Ph.D in

Power System Planning at Anna University, Chennai, Tamil Nadu. He has published three international journals and has four International/National conference publications. His research interest includes power system planning, voltage stability analysis and application of evolutionary algorithms to power system optimization.

M. SivaKumar received his B.E Degree in Electronics and communication Engg. From Dr. NGP. Institute of Technology, Coimbatore, Tamil Nadu in the year 2014, and pursuing M.E., in Applied Electronics, From Sree Venkateshwara Hi-Tech Engineering College, Gobi, TamilNadu.

DOI10.17148/IJARCCE.2017.63153