REAL-TIME HUMAN TRACKING BASED ON SWITCHING LINEAR DYNAMIC SYSTEM COMBINED WITH ADAPTIVE MEANSHIFT TRACKER

Zheyuan Li, Hong Liu, Chao Xu

Key Laboratory of Machine Perception and Intelligence
Peking University, Shenzhen Graduate School, P.R.China
Email: zheyuanli@cis.pku.edu.cn, hongliu@pku.edu.cn, xuchao@cis.pku.edu.cn

ABSTRACT

Real-time human tracking in complex environments usually presents many challenges, such as partial/complete occlusions caused by irregular motion and similar-color distractors. Switching Linear Dynamic System(SLDS) and Meanshift(MS) are two successful approaches, although both have inherent deficiencies, like accumulated prediction and correction errors in SLDS, uncoded attentional and spatial information in Meanshift, respectively. In this paper, a spatial-representive Meanshift and a joint attentional feature histogram from bottom-up and top-down attention models are used to make the tracker more adaptive. Then a tracking algorithm is proposed as adaptive Meanshift embedded SLDS, where four transition matrixes $A(s_t)$ handle partial/complete occlusions with irregular motion under the framework, and adaptive Meanshift solves short occlusions of similar-color distractors in local search for higher accuracy. Experiments show that this method can work more robustly for partial/complete occlusions among multiple persons compared with adaptive Meanshift and Meanshift embedded SLDS.

Index Terms— SLDS Model, Meanshift Tracker, Multi-cue Fusion, Human Attention Mechanism

1. INTRODUCTION

Real-time tracking is always a critical task in many applications such as autonomous robot navigation, intelligent surveillance, perceptual user interface, driver assistance and biotracking, etc [1, 2, 3]. And single target tracking remains a challenge due to frequent irregular motion, complete or partial occlusions and similar color disturbance between target and distractors.

There are probabilistic methods and deterministic methods to handle those problems. Probabilistic methods view tracking as a dynamic state estimation problem under probabilistic framework. Since human dynamics can provide powerful cues in occlusions, Dynamic Bayesian Network(DBN) models are usually used, such as Linear Dynamic System(LDS) and its derivatives[4, 5]. LDS is an optimal estimator based on linear prediction and minimized error covariances. However, it cannot model complex motion and tell precise location of target as a linear model. Like other DBNs, the accumulated prediction/correction errors require more accurate trackers. Meanwhile, deterministic methods compare target model with current frame and find out the most promising candidate, such as Meanshift and Trust Region[1, 6]. Meanshift is a kernel-based method climbing probabilistic density gradient to seek distribution mode and is more accurate than probabilistic methods. After it is applied by Comaniciu et.al. and Bradski, it cannot handle problems above very well, because it always employs single color cue which is vulnerable to complex environment.

At present, there are two solutions to solve these problems. One is to combine probabilistic methods with deterministic methods together for more accurate and robust algorithms, the other is to integrate multiple cues like motion, color and spatial information for enhancing original Meanshift. Zivkovic et.al. have analyzed and compared a range of approximate Bayesian schemes(LDS, mixed LDS, Partial Filtering, etc) with original meanshift especially on occlusions, finding they work efficiently[7]. Some researchers have established multi-cue integration and feature fusion mechanisms in target representive level[8, 9]. Others are working on higher-order spatiogram or tunable block weights to generate advanced tracker with surrounding information[10, 11].

In this paper, SLDS is employed as a tracking framework for two reasons: first, the Markov process in SLDS controls a LDS rather than a fixed Gaussian model; second, by mapping discrete hidden states to piecewise linear measurement models, SLDS has potentially greater descriptive power than a Hidden Markov Model in tracking[5, 12]. Besides, a spatial-representive Meanshift based on joint attentional feature histogram is built from color, motion information and two attention models. Therefore, SLDS combined with adaptive Meanshift is proposed, where SLDS predicts rough position when target moves irregularly and occlusion happens, and then the position is used to initiate Meanshift parameters. Thus, Meanshift can keep continuous and accurate tracking when target moves among distractors.

The rest of this paper is organized as follows: section 2 presents adaptive Meanshift and joint attentional feature histogram. The whole structure of algorithm is proposed in section 3. Section 4 shows experimental results and discussions. Finally, conclusions are drawn in section 5.

2. JOINT ATTENTIONAL FEATURE HISTOGRAM AND SPATIAL-REPRESENTIVE MEANSHIFT

As an appearance-based method, Meanshift employs iterations to find a candidate $p = \{p_u\}_{u=1}^m$ which is the most similar to the target model $q = \{q_u\}_{u=1}^m$ on single color(or other feature) distribution, with the similarity of two distributions based on Bhattacharyya coefficient[1]. Since single color histogram is vulnerable to partial/complete occlusions and similar color disturbance, a joint

This work is supported by National Natural Science Foundation of China (NSFC, No.60875050, 60675025), National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247), Shenzhen Scientific and Technological Plan and Basic Research program (No.JC200903160369A, JC201005280682A), Natural Science Foundation of Guangdong (No.9151806001000025).

probability distribution map integrating color and motion cues with bottom-up/bottom-down attention models is constructed. Then the map is projected into 1D joint feature histogram again for spatial-representive Meanshift search.

2.1. Joint Attentional Feature Histogram

First, a joint attentional distribution map $P_a(x, y)$ replaces original single color probability distribution map, for $P_a(x, y)$ can fuse multiple cues for object representation. In order to produce such map, there are two stages based on visual bottom-up and top-down attention models. Fused color and motion features and spatial fovea vision function are utilized to implement these two stages.

In bottom-up stage, HSV colorspace is used to handle illumination changes and color disruptors. The H channel is inaccurate with a low S near to 0, therefore, proper thresholds and weights among HSV are set to limit too low S and too high V values and to produce color probability distribution map $P_c(x, y)$ through back-projection.

Motion salient region is generated by horizontal/vertical projections of consecutive frame differences. Although target motion is complicated, displacements can still be obtained with appropriate threshold and affine transformations. After parameter estimation and compensation for affine motion between former and later frames, difference results are projected horizontally and vertically to avoid foreground aperture problem and to detect the most probable motion regions $P_m(x, y)$.

In the top-down stage, an exponential spatial attenuation function $F_{(x,y)}$ which simulates the foveal vision is applied to produce $P_a(x, y)$. It uses selective attention on predicted regions and suppresses peripheral distracters remarkably due to target model. Because when eyes are searching the most promising candidate in joint salient map $P_c(x, y) * P_m(x, y)$, attention always guides them into the most promising location where target may appear. Then the position becomes a fixation point(x, y) with the densest attention and that of the surround becomes less and less. The formula is written in equation1:

$$F_{(x,y)} = \exp\left(-\frac{\sqrt{(x-x_0)^2+(y-y_0)^2}}{c \cdot w \cdot h}\right) \tag{1}$$

where (x, y) is pixel coordinate in the frame, (x_0, y_0) is fixation point which is tracked result in last frame, c is an attenuated constant, w and h are frame width and height.

Hence, the joint attentional probability distribution map is $P_a(x, y) = P_c(x, y) * P_m(x, y) * F_{(x,y)}$. Then it generates the joint attentional feature histogram through a reversed "back-projection". Experiments demonstrate that it can enhance correct detection probability and accurate search effectively, especially in multiple human scenes.

2.2. Spatial-representive Meanshift Tracker

Since a pixel at the same location may contribute differently to candidate overtime due to motion and occlusions, spatial information should be encoded in object representation during search process as shown in[5]. Here the target is divided into M blocks($M = 9$), let $\{x_i\}_{i=1}^{n_h}$ denote n_h pixel locations of the candidate centered at y in current frame. Then the candidate representation becomes:

$$\hat{p}_u(y) = C_p \sum_{i=1}^{n_h} \delta[b(x_i) - u] \sum_{j=1}^{M} \phi_j^{(u)} k(\| \frac{x_i - y - \overline{z}_j}{h_j} \|^2) \tag{2}$$

where C_p means normalization constants, δ is the Kronecker delta function, $b(x_i)$ is the index of the histogram bin, $k(x)$ is the kernel profile, $\overline{z}_j$ denotes the center of block j, and h_j specifies the kernel range of decay for block j, respectively. The larger h_j is, the more slowly weights of pixels are decreasing, and vice visa.

Then ϕ_j is calculated to smooth the sudden change of the correlation of target model $F_j^* = \{f_{j1}^*, f_{j2}^*, \ldots, f_{jn}^*\}$ and candidate region $F_j = \{f_{j1}, f_{j2}, \ldots, f_{jn}\}$ for block j:

$$\phi_j = C_\phi \sqrt{1 - \left(1 - \frac{\sum_{i=1}^{n} f_{ij}^* f_{ij}}{\sqrt{\sum_{i=1}^{n} f_{ij}^*} \sqrt{\sum_{i=1}^{n} f_{ij}}}\right)^2} \tag{3}$$

where C_ϕ is a normalization constant and n is the bin number in histograms.

When target locates exactly on one block, the correlation value will be approximately to 1. When some occlusions happen, the block content may result in one small correlation value. Therefore, equation of location in next frame can incorporate correlation factors into weight functions to make tracker more adaptive. The meanshift vector is derived as follows:

$$y_{t+1} = \frac{\sum_{i=1}^{n_h} w_i \sum_{j=1}^{M} \phi_j g\left(\| \frac{x_i - y_t - \overline{z}_j}{h_j} \|\right)^2 \left(\frac{x_i - \overline{z}_j}{h_j^2}\right)}{\sum_{i=1}^{n_h} w_i \sum_{j=1}^{M} \phi_j g\left(\| \frac{x_i - y_t - \overline{z}_j}{h_j} \|\right)^2 \left(\frac{1}{h_j^2}\right)} \tag{4}$$

where $g(x) = -k'(x)$.

3. FRAMEWORK OF ADAPTIVE MEANSHIFT EMBEDDED SLDS

Due to accumulated prediction and correction errors through each update step, up to time t, total bias may be too huge to lose exact target location, especially in complex environments. Therefore, the adaptive Meanshift is incorporated into SLDS to ensure SLDS can predict and be updated independently with minor error.

Here SLDS doesn't model human complex gestures with high degrees of freedom, but estimates motion of target central point. Suppose that the whole process of target can be decomposed into 4 kinds of submotion: low-speed linear walk($\leq$ 1.4m/sec), state 1), high-speed linear walk(state 2), staying still(state 3) and turning around(state 4), and then SLDS consists of 4 correlated transition matrixes $A_{(s_t)}$. At first SLDS is constructed through initialization, and then joint attentional feature histogram is built for target representation and mode search. The target center is referred as initial value of meanshift for more accurate search. Then the search result is fed back as the measurement of SLDS for correction. Then in next frame, after SLDS finishes the rough search firstly, Meanshift follows as local search and this method will keep continuous tracking.

In SLDS, each LDS state/mode is associated with a dynamical process. It describes dynamics of complex nonlinear physical processes by switching among a set of linear dynamic models. SLDS can be expressed by state-space formula 5.

$$\begin{aligned} P_r(i, j) &= P_r(s_t = i | s_{t-1} = j) = \Pi(i, j) \\ X_{t+1} &= A(s_{t+1}) X_t + v_{t+1}(s_{t+1}) \\ Z_t &= C X_t + w_t \end{aligned} \tag{5}$$

where s denotes discrete states switching models with continuous hidden states $X_t \in \mathcal{R}^N$ and observations $Z_t \in \mathcal{R}^M$, and v_t is state noise, $v_t(s_t)) \sim N(0, Q(s_t))$. Similarly, w_t is measurement noise, $w_t \sim N(0, R)$. $A(s_t)$ and C are transition and observation matrixes, respectively.

Assume the LDS of target motion models a Gauss-Markov process with Gaussian noises. The switching model is a discrete first-order Markov process with state variables s_t. And it is defined with the state transition matrix Π and an initial state distribution π_0. The LDS and switching process are coupled with the dependence of $A(s_t)$ and Q on the switching state s_t: $A(s_t{=}i) = A_i, Q(s_t = i) = Q_i$. Therefore, joint distribution $P(Z_T, X_T, S_T)$ over variables of SLDS is:

$$P_r(S_0)\prod_{t=1}^{T-1}P_r(S_t|S_{t-1})P_r(X_0|S_0)\prod_{t=1}^{T-1}P_r(X_t|X_{t-1},S_t)\prod_{t=0}^{T-1}P_r(Z_t|X_t) \quad (6)$$

where Z_T, X_T and S_T denote sequences of the length T of observations and hidden states.

Besides, an affine image warping is used to model target motion between two consecutive frames. The state variable X_t is modeled as $[x, y, v_x, v_y, a_x, a_y]_t$, where $(x, y), (v_x, v_y), (a_x,\ a_y)$represent target position, velocity and acceleration, respectively. The observation vector is $Z_t = (x, y)$.

Learning period: Sample video training;
Initialization: Built the joint attentional probability distribution map $P_a(x, y)$ and feature histogram of target;
Initialize LDS state estimations and Meanshift tracking window location $\hat{X}_{0|-1,i}$, and $\Sigma_{0|-1}$;
Initialize maxmized posterior $J_{0,i}$;
Iteration:;
for *t=1:T-1* **do**
  for *i=1:S* **do**
    for *j=1:S* **do**
      Predict and filter LDS state estimates $\hat{X}_{t|t,i,j}$ and $Q_{t|t}, i, j$;
      Find transition probabilities $J_{t|t-1,i,j}$ from state i to j;
    end
    Find the best transition $A_{t-1,i}$ into state i;
    Update sequence probability $J_{t,i}$ and the LDS state estimates $\hat{X}_{t|t,i}$ and $Q_{t|t,i}$;
    Use adaptive Meanshift based candidate $P_a(x, y)$ and feature histogram for accurate search, and the result will be used as correction at stage to predict and filter LDS estimates and transition matrix;
  end
end
Find best final switching state i^*_{T-1} and backtrace the best switching sequence S^*_T;
Do RTS smoothing for $S = S^*_T$;

Algorithm 1: Algorithm of Adaptive Meanshift embedded SLDS

In learning step, generalized EM and other Bayesian learning methods can be used to find optimal values of parameters $A(s_t)$,$Q(s_t)4$,Pi,P_r. Since submotion patterns have been determined, X_0, $s_i = \{s_1, s_2, s_3, s_4\}$, $Q_(s_i)$, C, R, π_0 can be initialized by several video training, which include 3 videos of 2107 frames with same resolution and environments as test sequences. Assume that R is a constant matrix.

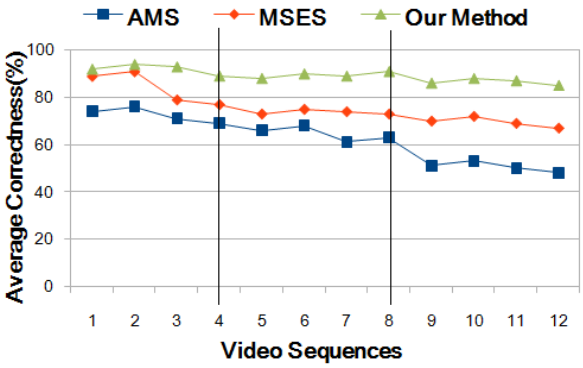

Fig. 1. Average Correctness of AMS, MSES and our method in 12 video sequences

In inference step, Viterbi approximate inference algorithm is used to find the most likely sequence of S^*_T for a given Z_T, [12], where the best switching sequence s^*_T equals to $\text{argmax}P_r(S^*_T|Z^*_T)$. Then the desired posterior $P(X_T, S_T|Z_T)$ is approximated by its mode $P_r(X_T|S^*_T, Z_T)$, and the transition probability from state i to j and from time $t-1$ to t is $J_{t|t-1,i,j}$. Our algorithm has shown its high robustness and efficiency to partial/complete occlusion caused by irregular motion and similar-color distrctors in human tracking.

4. EXPERIMENTS AND DISCUSSIONS

This proposed algorithm is tested on 12 real-scene sequences S1-S12 including 9017 frames. The tested video database includes various problems, for instance, 2 persons in cluttering environment with occlusions(S1- S4), 3 persons with different partial/complete occlusion and similar color disturbance(S5-S8), 5 persons with partial/complete occlusions, color distractors and irregular motions(S9-S12). All the sequences are captured at 15 frames per second with changing illumination in two $8 \times 8m^2$ halls. The resolution of each frame is 640×512 pixels and the original tracked region is 30×20 pixels. In this paper, the Adaptive Meanshift(AMS), Meanshift Embedded SLDS(MSES) and our method have been accomplished to illustrate their performances.

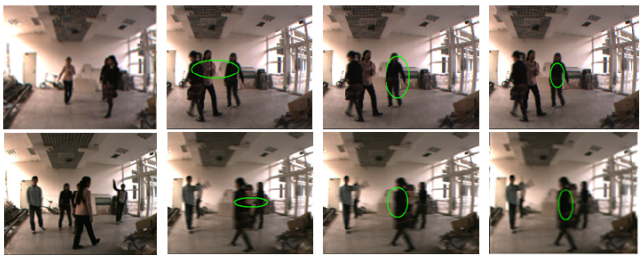

Fig. 2. S-10/11: tracking results of AMS/MSES/our algorithm with partial or complete occlusions. Both at frames #107, #152, #152, #152

It can be seen that our method and MSES both perform better than AMS in all sequences in Fig.1, especially in the complex environments of S9-12. For the switching Bayesian model can keep target dynamics and bring in robustness in whole process. And our

method shows higher accuracy than both MSES and AMS in sequences with occlusions, color disturbance and irregular motion together in S9-12. Because the adaptive meanshift tracker changes kernel weights adaptively and SLDS switches according to current motion pattern more accurately. Meanwhile, in S1- S4 where occlusions occur between 2 persons without similar color disturbance, both AMS and MSES show same experimental trends because the spatial information doesn't provide essential clues in search. And our method has a better result than others in S5-S8 because there are multiple color distractors exist in S5-S6 and irregular motion happens in S7-S8 among 3 persons.

In Fig.2 there are two sequences where complete and partial occlusions occurred. The upper row from S-10 shows the girl with dark red coat and black trousers is our target. And she is being occluded completely by the girl wearing dark red coat and plaid skirt at frame 107. Then at frame 152 we see three kinds of results using AMS, MSES and our method. In the second picture, AMS tried to distinguish the distractor and target region after occlusion, but finally fails. The MSES and our method both followed the right person, however, MSEKF didn't adjust the search window scale properly and produces some residuals.

Fig. 3. S-12: tracking for occlusions caused by irregular motions: 1st to 3rd rows: MSES; 4th to 6th rows: our method. Both at frames #113, #424, #430, #437, #438, #439, #445, #451, #457,#463, #475, #481, #487, #493, #505

In Fig.3 there are five persons and target is the girl in a dark red coat and plaid skirt. Because of their complex motion, continuous occlusions, like at frame 437, our method can predict location of the most likely region by using proper LDS model and meanshift will use non-occluded body part information and attention models adaptively to find the location mode. However, the MSES tracker can be disturbed by cluttered clues such as the boy with black trousers and another girl with dark suit, resulting larger search region and wrong location, like frames after 439, because it can just choose proper LDS in certain motion-patterns without the combined surrounding information like spatial, color cues.

5. CONCLUSIONS

This paper proposes a novel tracking algorithm combined with SLDS and adaptive Meanshift tracker. The SLDS is integrated to predict rough position, and to handle partial/complete occlusions with irregular motion, then adaptive meanshift searches precise information for solving occlusions caused by similar-color distractors and model update. Moreover, spatial-representive Meanshift is generated from joint attentional distribution histogram and multi-cue strategy. Extensive experiments show that this method performs very well under some challenging situations with partial/complete occlusions compared with the adaptive Meanshift and Meanshift Embedded SLDS.

6. REFERENCES

[1] D.Comaniciu,V.Ramesh and P.Meer, "Kernel-Based Object Tracking", *IEEE Trans. on Pattern Analysis and Machine Intellengece*, vol.25(5), pp.564-577, 2003.

[2] G.R. Bradski, "Computer vision face tracking for use in a perceptual user interface", *IEEE Workshop on Applications of Computer Vision*, pp.214-219,1998.

[3] D.Comaniciu, V.Ramesh and P.Meer, "Real-time tracking of Non-rigid Object Using Mean-shift", *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.142-149, 2000.

[4] G.Welch and G.Bishop, "An Introduction to the Kalman Filter", *SIGGRAPH, In Computer Graphics, Annual Conference on Computer Graphics and Interactive Techniques*, ACM Press, Addison-Wesley, Los Angeles, CA, USA SIGGRAPH, 2001.

[5] E.B. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky, Nonparametric Bayesian Learning of Switching Dynamical Systems, *NIPS*, Vancouver, Canada, 2008.

[6] T.-L. Liu and H.-T. Chen, "Real-time tracking using trust-region methods"*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.26(3), pp.397 - 402, 2004.

[7] Zoran Zivkovic, Ali Taylan Cemgil and Ben Krosem, "Approximate bayesian methods for kernel-based object tracking", *Computer Vision and Image Understanding(CVIU)*, pp.743-749, 2009.

[8] M.Spengler and B.Schiele, "Toward robust multi-cue intrgration for visual tracking", *Machine Vision and Applications*, Vol.14, pp.50-58, 2003.

[9] Y.Wu and T.S.Huang, "A Co-inference Approach to Robust Visual Tracking",*Proc. of Int. Conference on Computer Vision*, Vol.2, pp.26-33, 2001.

[10] Ying Shi, Hong Liu, Yi Liu and Hongbin Zha, "Adaptive Feature-spatial Representation for Mean-shift Tracker", *Proc.IEEE International Conference on Image Processing*, pp.217-220, 2007.

[11] Stanley T,Birchfield and Sriram Rangarajan, "Spatialgrams versus histograms for region-based tracking", *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[12] Vladimir Pavlovic, James M. Rehg, and John Maccormick, "Learning Switching Linear Models of Human Motion",*NIPS*, pp.981-987, 2000.