COUPLED HIDDEN MARKOV MODELS FOR ROBUST EO/IR TARGET TRACKING

Jiading Gai, Yong Li, Robert L. Stevenson

Department of Electrical Engineering, University of Notre Dame
275 Fitzpatrick Hall, Notre Dame, IN 46556
{jgai, yli5, rls@nd.edu}

ABSTRACT

Augmenting Electro-Optical (EO) based target tracking systems with Infrared (IR) modality has been shown to be effective in increasing the accuracy rate of the tracking system. A key issue in designing such a multimodal tracking system is how to combine information observed from different sensor types in a systematic way to obtain desirable performance. In this paper, we present an investigation into integrating EO and IR sensors within Hidden Markov Model (HMM) based frameworks. We propose to use a Coupled Hidden Markov Model (CHMM) to improve upon the existing fusion schemes. Another contribution is that we propose to use a robust t-distribution based subspace representation in the CHMM to model appearance changes of the target. Numerical experiments demonstrate that the proposed CHMM tracking system has improved performance over other integration schemes for situations where the target object is corrupted by noise or occlusion.

Index Terms— Coupled hidden markov model, target tracking, subspace representation, t-distribution, CONDENSATION algorithm

1. INTRODUCTION

Force protection, intelligence gathering, and targeting systems all use sensors to track interesting targets and collect necessary information about the environment so that intelligent decisions can be made about proper course of action. In most systems a single sensor type (i.e., EO) is utilized. While the visible spectrum is a major source of target information, the infrared spectrum is considered to be a useful supplementary source for providing a better sensing in various conditions [1].

Single sensor based target tracking systems work well for many applications, but their performance degrades considerably when the EO sensor has the similar response to the target and to the background (for example, when the target object wears a camouflage suit.) Limited success using a single sensing technology for object tracking opens the questions of how to utilize infrared sensor effectively to complete the task. Several methods have been proposed on the fusion of EO/IR sensors for object tracking. Conaire et al. [2] combines visible and infrared spectrum data to maintain a nonparametric background model. The target region in each frame is identified by comparing the current frame with the background. This algorithm transforms an object tracking problem into an object detection problem and fails to incorporate temporal information. Goubet et al. [8] provides an update to up survey on this line of research. Kang et al. [1] formulates EO/IR tracking as a joint maximization of a set of probability models using joint probability data association filter. Since the authors do not assume the Markov property for the hidden states, the algorithm has high computational load. Leykin et al. [3] formulates the multimodal tracking problem under the well-known HMM. HMM provides a powerful framework to incorporate temporal information and multiple cues. The authors use a combined color input observed from EO and IR sensors as features for tracking. The fusion scheme forms composite EO/IR feature vectors by simply concatenating the vectors from each sensor and the multimodal tracking is performed in the combined feature space. One disadvantage of integrating EO/IR sensors using the plain HMM is that the system becomes very sensitive to sensor alignment error.

Based on the analysis above, we propose the use of coupled hidden Markov models (CHMM) to model the interaction between the EO and IR modes through time (See Fig. 1). The CHMM was first introduced by Brand [6] to capture the interprocess influences across time in the systems of multiple interacting processes. The CHMM brings in the robustness against sensor alignment error by treating each sensor as a subsystem. The two subsystems are then coupled by bridging the states of both sensors, i.e., each hidden state at time t has a transition distribution parameterized by the two hidden states from both the EO and IR sensors at time $t-1$. This offers an ideal framework for closely coupling of the EO and IR sensors.

Another distinguishing point of our work is that we propose to use a robust t-distribution based eigenspace representation to reflect the appearance changes of the target and to model the observation distribution in the CHMM. In our previous study [4], we show how replacing a Gaussian distribution based subspace representation (i.e., probabilistic principal component analysis) with a robust t-distribution based

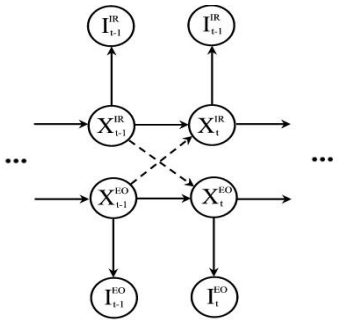

Fig. 1. A grahical model for the proposed CHMM.

subspace representation can help achieve a better outlier rejection mechanism and improve robustness of the unimodal tracking systems. Here, we adopt this robust subspace representation into our multimodal tracking framework. Another reason for selecting robust eigenbasis as features for tracking is that, arguably, dense features are more robust for object tracking in non-ideal situations than sparse features such as points, edges and contours, since in dense features every location within the target region gets to contribute to the final decision, as contrast to the sparse features where only a small portion of pixels are exploited.

2. EO/IR TARGET TRACKING WITHIN THE CHMM FRAMEWORK

We assume that the data from the two sensors has been synchronized in time properly prior to tracking. Each modality is modeled with an HMM. The hidden states $\{X_t^{EO}, X_t^{IR}\}$ describe the motion and position parameters of the target in the EO and IR sensors respectively at time t. The states of the two HMMs are coupled in time through conditional probabilities between the hidden state variables $p(X_t^{EO}|X_{t-1}^{EO}, X_{t-1}^{IR})$ and $p(X_t^{IR}|X_{t-1}^{EO}, X_{t-1}^{IR})$. $\{I_t^{EO}, I_t^{IR}\}$ represent the target appearances observed from the two modalities.

Given a set of the observed target images $\{I_i^{EO}, I_i^{IR}\}_{i=1}^t$ from the two sensors, we recursively update the posterior distribution $p(X_t^{EO}, X_t^{IR}|\{I_i^{EO}, I_i^{IR}\}_{i=1}^t)$ over the state variables X_t^{EO}, X_t^{IR} as follows:

$$p(X_t^{EO}, X_t^{IR}|\{I_i^{EO}, I_i^{IR}\}_{i=1}^t) \propto p(I_t^{EO}, I_t^{IR}|X_t^{EO}, X_t^{IR}) \int p(X_t^{EO}, X_t^{IR}|X_{t-1}^{EO}, X_{t-1}^{IR}) p(X_t^{EO}, X_t^{IR}|\{I_i^{EO}, I_i^{IR}\}_{i=1}^{t-1}) dX_{t-1}^{EO} dX_{t-1}^{IR}. \quad (1)$$

According to the dependence structure encoded in the graphical model of our CHMM framework, the observation model $p(I_t^{EO}, I_t^{IR}|X_t^{EO}, X_t^{IR})$ can be written as:

$$p(I_t^{EO}, I_t^{IR}|X_t^{EO}, X_t^{IR}) = p(I_t^{EO}|X_t^{EO}) p(I_t^{IR}|X_t^{IR}). \quad (2)$$

The tracking process is also governed by the motion model $p(X_t^{EO}, X_t^{IR}|X_{t-1}^{EO}, X_{t-1}^{IR})$, which predicts the hidden states at time t given the previous states. The motion model can also be factorized accordingly as:

$$p(X_t^{EO}, X_t^{IR}|X_{t-1}^{EO}, X_{t-1}^{IR}) = p(X_t^{EO}|X_{t-1}^{EO}, X_{t-1}^{IR}) \cdot p(X_t^{IR}|X_{t-1}^{EO}, X_{t-1}^{IR}). \quad (3)$$

The dynamics of motion of the moving object is approximated by an affine image warping [7] and the state variable $X_t = (x_t, y_t, \theta_t, s_t, \alpha_t, \phi_t)$ describes the translation, rotation angle, scale, aspect ratio and skew direction of the target motion at time t in each sensor. The motion models are set to be a mixture of Gaussian distribution parameterized by X_t^{EO}, X_t^{IR}:

$$p(X_t^{EO}|X_{t-1}^{EO}, X_{t-1}^{IR}) = \lambda \cdot N(X_t^{EO}; X_{t-1}^{EO}, \Psi^{EO}) + (1-\lambda) \cdot N(X_t^{EO}; X_{t-1}^{IR}, \Psi^{IR}),$$
$$p(X_t^{IR}|X_{t-1}^{EO}, X_{t-1}^{IR}) = \lambda \cdot N(X_t^{IR}; X_{t-1}^{IR}, \Psi^{IR}) + (1-\lambda) \cdot N(X_t^{IR}; X_{t-1}^{EO}, \Psi^{EO}).$$

where Ψ^{EO}, Ψ^{IR} are taken to be diagonal covariance matrices to reduce the number of parameters that need to be estimated. $N(\cdot)$ denotes the Gaussian density function. λ is set to 0.6 in our simulation.

The construction of the observation model merits a detailed explanation. Since each modality maintains its own observation distribution of the same mathematical form, we choose not to highlight sensor types in the derivation below. As we pointed out in our previous study [4], classic gaussian based eigenspace representation (a.k.a. probabilistic principal component analysis) has robustness issues with respect to atypical observations. Based on the fact that true data with numerous outliers exhibit "heavy tails", we proposed to use a robust t-distribution based eigenspace representation instead. The t-distribution has a heavier tail than Gaussian. The thickness of its tail is regulated by the degree of freedom parameter v. Gaussian distribution is a special case of the t-distribution when v goes to infinity. So the robust eigenspace representation is a generalized version of the classic one.

The latent variable model that we use to describe the robust eigenspace representation is:

$$I_t = WY_t + \mu + \epsilon_n, \quad (4)$$

where I_t is the observed target image of size $D \times 1$ in raster-scan format; W is a $D \times d$ projection matrix that relates I_t and Y_t. The columns of W encode the first d principal components of the image set of the target appearance up to the time t; Y_t is a latent variable; μ is the mean and ϵ_n is the noise term. Y_t is the decomposition coefficients of the target image (determined by X_t) in the current robust eigenbasis. To simplify our derivation, we use an auxiliary random

variable $w_t \sim \Gamma(\frac{v}{2}, \frac{v}{2})$ and define the following conditional probabilities:

$$p(Y_t|w_t) = N(0, \frac{I_d}{w_t}), \quad (5)$$

$$p(I_t|Y_t, w_t) = N(WY_t + \mu, \frac{\sum}{w_t}), \quad (6)$$

where I_d is a d-dimensional identity matrix, $\sum$ is a diagonal covariance matrix.

Using (4)-(6), the observation model $p(I_t|X_t)$ is computed through proper integration [9, 10]:

$$p(I_t|X_t) = \frac{\Gamma(\frac{v+D}{2})|WW^{'} + \sum|^{-1/2}(v\pi)^{-D/2}}{\Gamma(\frac{v}{2})} \cdot \left(\frac{(t-\mu)^{'}(WW^{'} + \sum)^{-1}(t-\mu)}{v} + 1\right)^{(-\frac{v+D}{2})}, \quad (7)$$

where $(\cdot)'$ is the matrix transpose.

Note that $p(I_t|X_t)$ is a t-distribution with mean μ, covariance matrix $WW^{'} + \sum$ and the degree of freedom v.

3. FITTING THE PROPOSED CHMM WITH EM ALGORITHM

Each iteration of the proposed EO/IR tracking algorithm operates in two stages, starting with estimation of the position and appearance of the target at the current frame and concluding with parameters update for the robust subspace representation. In this section, we focus on how to estimate a set of values for the CHMM parameters that best represents the set of tracked target images up to the current frame. The model parameters are $\{\mu, v, \sum, W\}$. The maximum likelihood estimation (MLE) of $\{\mu, v, \sum, W\}$ is obtained using Expectation-Maximization (EM) algorithm by maximizing the expected complete data log-likelihood [11, 4].

The expectation quantities needed to compute the new set of parameters in the M-step are given by:

$$\overline{w}_t = \frac{v + D}{v + (I_t - \mu)^{'} M (I_t - \mu)}, \quad (8)$$

$$\overline{w_t Y_t} = (W^{'}\Sigma W + I_d)^{-1} W^{'} \Sigma^{-1} (I_t - \mu)\overline{w}_t, \quad (9)$$

$$\overline{w_t Y_t Y_t^{'}} = (W^{'}\Sigma W + I_d)^{-1} + \overline{w}_t \overline{w_t Y_t} (\overline{w_t Y_t})^{'}, \quad (10)$$

$$\overline{\log w_t} = \frac{d\Gamma(\alpha)}{d\alpha}|_{\alpha=\frac{v+D}{2}} \cdot \frac{1}{\Gamma(\frac{v+D}{2})} - \log \frac{v + (I_t - \mu)^{'} M (I_t - \mu)}{2}, \quad (11)$$

where $\overline{\star} = \mathrm{E}_{Y_t, w_t | I_t, \{\mu, v, \sum, W\}}[\star]$, $M = (WW^{'} + \Sigma)^{-1}$.

In the M-step, the expected complete log likelihood is maximized with respect to $\{\mu, v, \sum, W\}$. This gives the update rules for the model parameters:

$$\mu = \frac{\sum_{t=1}^{T} \overline{w}_t I_t - \sum_{t=1}^{T} W\overline{(w_t Y_t)}}{\sum_{t=1}^{T} \overline{w_t}}, \quad (12)$$

$$W = (\sum_{t=1}^{T} \Sigma^{-1}(I_t - \mu)\overline{w_t Y_t^{'}})(\sum_{t=1}^{T} \overline{Y_t Y_t^{'} w_t})^{-1}, \quad (13)$$

$$\Sigma = \frac{1}{T}\sum_{t=1}^{T} \mathrm{diag}\{(I_t - \mu)(I_t - \mu)^{'}\overline{w_t} - 2W\overline{w_t Y_t}(I_t - \mu)^{'} + W\overline{(Y_t Y_t^{'} w_t)}W^{'}\}, \quad (14)$$

$$v = \arg\max_v \{\frac{Nv}{2}\log\frac{v}{2} - N\log\Gamma(\frac{v}{2}) + (v-2)\sum_{t=1}^{T}\overline{\log w_t} - \sum_{t=1}^{T} v\overline{w_t}\}, \quad (15)$$

where T denotes the index of the current frame, diag(A) denotes the diagonal matrix consisting of the diagonal elements of A.

Note that, in practice, we need to bound the maximization of v in (15). v has to be greater than zero and less than certain maximum degrees of freedom whose value depends on the data.

4. INFERENCE IN THE CHMM

We use the CONDENSATION algorithm to perform an approximate inference in the CHMM. The CONDENSATION algorithm [12] is a well-known sequential Monte Carlo method for approximating Bayesian inference, where the posterior probability is recursively approximated with a randomly generated set of weighted samples, called particles. We use 300 particles sampled from $p(X_t^{EO}, X_t^{IR}|X_{t-1}^{EO}, X_{t-1}^{IR})$ and weighed by $p(I_t^{EO}, I_t^{IR}|X_t^{EO}, X_t^{IR})$ to approximate the posterior probability $p(X_t^{EO}, X_t^{IR}|I_t^{EO}, I_t^{IR})$. A new set of particles is generated from the current set by random sampling proportionally to these weights. The optimal solution for $\{X_t^{EO}, X_t^{IR}\}$ is approximated by the strongest mode of the particle distribution.

5. RESULTS

In this section, we report results that demonstrate the superiority of the proposed CHMM framework over the existing fusion schemes for EO/IR tracking.

The data was collected by a pair of fixed cameras at infrared and visible wavelengths respectively. The sensors were roughly aimed at the same scene. The multimodal data sequences were pre-registered using homography with invariant features across sensors. In the video sequences, a person

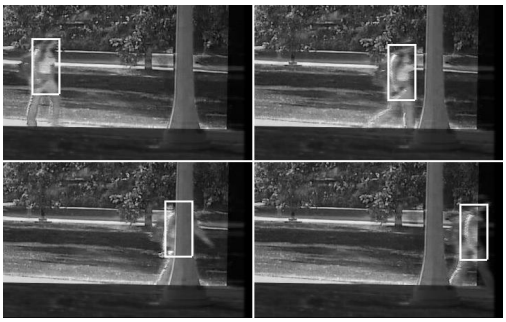

Fig. 2. EO/IR tracking w.r.t partial occlusion. Tracking results are shown from left to right, top to bottom.

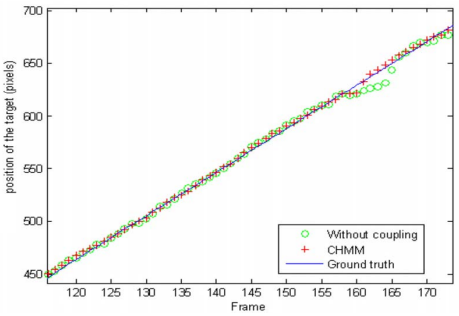

Fig. 3. Tracking acurracy compared to ground truth.

is getting partially occluded by a street lamp as time goes by. The target region reappears after the occlusion is passed. In Fig. 2, experiment illustrates that our CHMM scheme produces robust tracking results despite occlusion. In Fig. 3, using the same sequence, we compared the results from the proposed CHMM based tracker against the fusion scheme that ignores the states coupling between the two sensors (i.e., the scheme that treats the EO and IR sensors as two independent hidden Markov processes. The fusion only takes place at the decision level.) Since the target is travelling at a constant speed, the ground truth positions can be easily marked and also included in the comparison. The occlusion begins at the 158^{th} frame and ends at the 167^{th} frame.

6. CONCLUSION

In this paper, we proposed a new CHMM framework targeted at closely coupled EO/IR multimodal tracking. The image sequences acquired by the two sensing modes are modeled as two hidden markov processes. The states of the different HMMs are bridged together to form a CHMM. To ensure the robust performance of our system, we further develop a t-distribution based subspace representation to cope with various outliers. Experiments on real-world sequences show an obvious improvement in terms of tracking accuracy.

7. REFERENCES

[1] J. Kang, K. Gajera, I. Cohen, and G. Medioni, "Detection and Tracking of Moving Objects from Overlapping EO and IR Sensors", Proceedings of the Joint IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum (OTCBVS'04), Washington, DC, July, 2004.

[2] C. Conaire, E. Cooke, N. O'Connor, N. Murphy, A. Smeaton, "Background Modelling in Infrared and Visible Spectrum Video for People Tracking", International Conference on Computer Vision and Pattern Recognition, San Diego, CA, June20-25 2005.

[3] A. Leykin, R. Hammoud, "Robust Multi-Pedestrian Tracking in Thermal-Visible Surveillance Videos", Computer Vision and Pattern Recognition Workshop, 2006 Conference on June 17-22 2006 Page(s):136 - 136.

[4] J. Gai, R.L. Stevenson, "A robustified Hidden Markov Model for visual tracking with subspace representation", Visual Communications and Image Processing 2007(VCIP 2007), San Jose, California, USA, Jan 28-Feb 01, 2007.

[5] C. Schmid, and R. Mohr, "Local grayvalue invariants for image retrieval", IEEE PAMI, 19, 5 (1997), pp. 530-534.

[6] M. Brand, "Coupled hidden Markov models for modeling interacting processes", MIT Media Lab Perceptual Computing/Learning and Common Sense Techincal Report 405 (Revised), June 1997.

[7] J. Lim, D. Ross, R. Lin, M. Yang, "Incremental Learning for Visual Tracking," in *Advances in Neural Information Processing Systems 17, MIT Press*, 2004.

[8] E. Goubet, J. Katz, F. Porikli, "Pedestrian Tracking Using Thermal Infrared Imaging," SPIE Conference Infrared Technology and Applications XXXII, Vol. 6206, pp. 797-808, June 2006

[9] Z. Khan and F. Dellaert, "Robust Generative Subspace Modeling: The Subspace t Distribution," in *Georgia Institute of Technology GVU Center, GIT-GVU-04-11*, 2004.

[10] C. Archambeau, N. Delannay and M. Verleysen, "Robust Probabilistic Projections," in *Machine Learning, International Conference on*, 2006.

[11] G.J. McLachlan and T. Krishman, "The EM Algorithm and extensions," *Wiley*, 1997.

[12] M. Isard and A. Blake "CONDENSATION – conditional density propagation for visual tracking," Int. J. Computer Vision, 29, 1, 5–28, (1998)