MODEL-BASED TRACKING: TEMPORAL CONDITIONAL RANDOM FIELDS

M.J. Shafiee[†], *Z. Azimifar*[†], *P. Fieguth*[‡]

[†]School of Electrical & Computer Engineering, Shiraz University, Shiraz, Iran
{mj_shafiee, azimifar}@cse.shirazu.ac.ir
[‡]Systems Design Engineering, University of Waterloo, Waterloo, Canada
pfieguth@uwaterloo.ca

ABSTRACT

We present *Temporal Conditional Random Fields*, a probabilistic framework for modeling object motion. The state-of-the-art discriminative approach for tracking is known as dynamic conditional random fields. This method models an event based on spatial and temporal relation between pixels in an image sequence without any prediction. To facilitate such a powerful graphical model with prediction and come up with a CRF-based predictor, we propose a set of new temporal relations for object tracking, with feature functions such as optical flow (calculated among consequent frames) and line filed features. We validate our proposed method with real data sequences and will show that the TCRF prediction is nearly equivalent with result of template matching. Experimental results indicate that our TCRF can predict future state of any maneuvering target with nearly zero error during its constant motion. Not only the proposed TCRF has a simple and easy to implement structure, but also it outperforms the state-of-the-art predictors such as Kalman filter.

***Index Terms*—** Visual Tracking, Discriminative Models, Conditional Random Fields, Potential Function

1. INTRODUCTION

Event modeling has generated a large body of research during the past two decades. In modeling real events we usually have somehow corrupted or insufficient measurements, making the problems ill-posed. In such cases some sort of prior knowledge or constraints may be applied to allow a problem solution. Object tracking, image denoising and surface reconstruction are among the problems that are addressed by statistical modeling.

Our objective is to model the target dynamic in object tracking problems. A significant number of methods have been proposed to solve this problem, some statistical and some heuristic, where the Kalman filter is the optimum least-squares framework in the presence of Gaussian measurement noise. The Kalman filter, and its many variations, predicts the next state of the object with a predefined dynamic and then updates the predicted state with measurements.

Within the context of graphical modeling, the Kalman filter behavior is like that of Hidden Morkov Models (HHM), based on a generative model that models the joint distribution of measurements and label. Similarly, in computer vision, Markov Random Fields (MRF) are another generative model applied, assuming conditional independency between measurements when conditioned on labels (states).

Motivated by the modeling success of MRFs, Conditional Random Fields (CRF) were introduced, directly modeling the conditional probability distribution of labels given measurements, and relaxing the conditional independence assumption. Whereas many image processing problems have limited data for the purpose of modeling, leading to comparatively inaccurate MRF models, the key idea is the CRF can better solve many problems of computer vision because it explicitly models only the conditional distribution, and does not attempt to learn a prior.

In this paper we propose a new probabilistic approach to object tracking based on the CRF. In a discriminative framework, the object motion using a number of frames is modeled by a Temporal Conditional Random Field (TCRF). After training the TCRF and finding weights corresponding to each potential function, we use the TCRF to predict the new state of the target.

2. TRACKING

Taycher *et al.* [1] proposed human tracking based on a conditional random field, with an L_1 similarity space corresponding as the potential functions. In this work different poses were considered as states for tracking within a sequence of images, where the number of states was defined ahead of time.

CRFs are also applied to image-sequence segmentation [2, 3], where the random fields are modeled using spatial and temporal dependencies.

Sigal *et al.* [4] used two-layer spatio-temporal models for component-based detection and tracking of objects in video sequences. Each object or component of an object is considered as a node of a graphical model at a given time. Moreover, the graph edges correspond to learned spatial and temporal constraints. Following this work, Ablavsky *et al.* [5] proposed a layered graphical model for partially occluded object tracking. A layered image-plane represents motion around a known object that is associated with a pre-computed graphical model.

978-1-4244-7994-8/10/$26.00 ©2010 IEEE

All of the approaches discussed above suffer from complexity and required pre-processing stages. Furthermore these works were proposed for specific purposes and are not associated with any prediction phase. In this paper we aim to model the object motion in a simple and general way. It is assumed that there is no sudden change in position and motion direction of the object along the consecutive frames. Initially, the motion conditional distribution $P(Y|M)$ is modeled by TCRF using the first two frames where measurements are shown as M and prediction states as Y. Then, the object position in the next frame can be predicted given the current frame. We empirically observed that, often, the prediction has no error when no change occurs in the object motion. After the prediction stage a heuristic procedure searches around the predicted coordinates to find the best matching sub-image with a target template extracted using the two last training frames. If the TCRF prediction for time t has significant difference with matching coordinates of the template, the CRF is again trained with frames $t + 1$ and t.

3. TEMPORAL CONDITIONAL RANDOM FIELDS

The idea of a conditional random field was first proposed by Lafferty *et al.* [6]. It is a discriminative model that relaxes the conditional independence assumption of generative models by directly estimating the conditional probability of labels given measurements. The general form of a CRF is

$$P(Y|M) = \frac{1}{Z(M)} \quad \prod_{c\in C} \prod_{\phi_c \in c} \exp\Big\{ \sum_k \lambda_{\phi_c,k} f_{\phi_c,k}(Y_{\phi_c}, M) \Big\} \tag{1}$$

where $Z(M)$ is a constant normalization with respect to all possible values of Y, C represents the set of clique templates, $f_{\phi_c}(Y_{\phi_c}, M)$ is a potential function with respect to clique ϕ_c, and finally λ shows the weight of each potential function.

Early CRFs use only spatial relations among random fields, thus, Sutton *et al.* [7] proposed dynamic conditional random fields (DCRF) to capture spatial relations between neighbor nodes and temporal relations across temporally-separated frames [8]. The DCRF introduces two new terms to the original definition of CRF, a single temporal feature function and an interaction temporal feature function.

In this paper, we study temporal conditional random fields with feature functions describing temporal relations between successive frames. In other words, our objective is to investigate the adding of *prediction* to the CRF, in which case one frame is considered as a measurement of the next frame. For modeling object motion, a simple temporal relation among frames is sufficient, since we assume that the tracked object is rigid. Eq.(2) shows the formal definition of our TCRF:

$$\begin{aligned} P(Y_{t+1}|m_t) \quad &= \quad \frac{1}{Z(m_t)} \exp\Big\{ \sum_{y_i \in Y_{t+1}} \Big\{ \sum_{k_1} \lambda_{k_1} f_{k_1}(y_{t+1,i}, m_t) \\ &+ \quad \sum_{k_2} \lambda_{k_2} f_{k_2}(y_{t+1,i}, y_{t+1,N_i}, m_t) \Big\}\Big\} \end{aligned} \tag{2}$$

where $y_{t+1,j}$ labels any pixel j within frame $t + 1$ as foreground or background and m_t is an observation of frame t. In our temporal CRF we use two kinds of feature functions $f(y_{t+1,i}, m_t)$ and $f(y_{t+1,i}, y_{t+1,N_i}, m_t)$ as single potential and interaction potential functions, respectively, and N_i is the set of neighbors for each node i.

Our goal is to study the effect of different potential functions in modeling object motion in the context of CRFs. Because tracking problems are inherently temporal, clearly our potential functions need to be based on motion features with some sort of temporal dependency. The following sections summarize the functions that we propose to use.

3.1. Optical Flow

Optical flow is an approximation of motion based upon local derivatives in a given sequence of images [9]. Optical flow estimates pixel movement adjacent images, based on the intensity invariance assumption that

$$I(x, y, t) = I((x + \delta x), (y + \delta y), (t + \delta t)) \tag{3}$$

which can be rewritten, after a first-order Taylor expansion and simplification, as

$$(I_x, I_y).(v_x, v_y) = -I_t \tag{4}$$

where (I_x, I_y) is the spatial intensity gradient and v_x and v_y are motion velocities in the x and y directions.

3.2. Line Field

Line fields were first introduced by Geman and Geman [10], as a hidden binary model indicating the presence (state = 1) or absence (state = 0) of edges. Here, we define a slightly different form of the original definition in [10]:

$$F(m_t, Y_{t+1,N_i}, i) = \sum_{j\in N_i} 1 - \delta(m_t(i) - y_{t+1}(j)) \tag{5}$$

where $\delta(a)$ is Kronecker delta function. We also exploit the duality of feature functions in order to reduce the similarity between the value function of each configuration in temporal relation neighbors and to reinforce feature functions be more discriminative. The dual of (5) is:

$$F(m_{t,\bar{N}_i}, Y_{t+1}, i) = \sum_{j\in \bar{N}_i} 1 - \delta(y_{t+1}(i) - m_t(j)) \tag{6}$$

where N_i and $\bar{N}_i$ are the neighborhood sets of i in fields Y and m, respectively.

3.3. Ising

The Ising model [10] is a classic, very simple binary prior model. The CRF does not require a prior model, however the local, four-neighbor Ising model can be adapted and modified into potential form as

$$F(m_t, Y_{t+1,N_i}, i) = \sum_{j\in N_i} m_t(i) \times y_{t+1}(j) \tag{7}$$

with the following dual:

$$F(m_{t,\bar{N}_i}, Y_{t+1}, i) = \sum_{j\in \bar{N}_i} m_t(j) \times y_{t+1}(i) \tag{8}$$

3.4. Consistency Measure

Yin [8] introduced a feature function to evaluate the consistency between two adjacent hidden nodes, such that when a pixel has state s_i at some time t, this state value is likely to spread to its neighbors in the following time steps:

$$
\begin{aligned}
F(m_t, Y_{t+1,N_i}, i) &= \sum_{j \in N_i} \delta(m_t(i) - y_{t+1}(j)) \\
F(m_{t,\bar{N}_i}, Y_{t+1}, i) &= \sum_{j \in \bar{N}_i} \delta(m_t(j) - y_{t+1}(i)) \quad (9)
\end{aligned}
$$

4. TRACKING MODEL

We recognize that object tracking has a vast literature, spanning many years. The purpose of this paper is not to contest that literature, or to make claims of superior tracking, rather we wish to study the effectiveness and the potential of CRF methods in tracking, specifically in prediction. Our goal is to study several candidate potential functions, and to assess their ability in the context of CRF prediction.

Given a temporal potential function, we can create a temporal CRF to predict object position in future frames. After predicting the position of an object, a heuristic method (such as template matching) searches around the predicted position to find the coordinates of the best matched candidate. If these predicted and matched coordinates are very different, then the CRF training is repeated using two last frames.

4.1. MAP Predicion

After training the TCRF, we use maximum a posteriori (MAP) estimation to predict the object position at time $t+1$. Since the TCRF models the object motion, prediction is accomplished by evaluating the probability of the next target location around its previous position at time t. The object's predicted position in frame $t+1$ is obtained from an ensemble of frames, such that we create a set of synthetic images with a synthesized target obeying a variety of dynamics, with the probability of each image assessed by the TCRF. The predicted target position (in frame $t+1$) is found from that sample with maximum probability.

4.2. Position Update

A separate update step serves to validate the TCRF. A template matching procedure is used to obtain the exact position of the object in each scene, where the template is constructed using the measurement and label frames that are used for training the TCRF. Experiments show that the predicted positions and template matching coordinates are often equivalent. If these two coordinates are different, the TCRF must be re-trained with last two frames.

5. RESULTS AND DISCUSSION

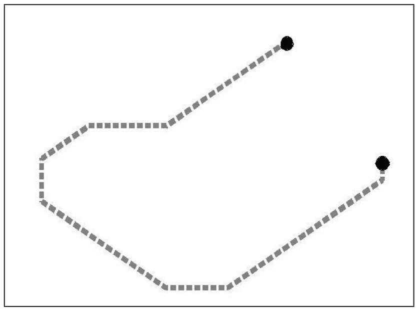

Figure 1: A sample of the maneuvering target used to test the proposed method. The object motion starts at top right of the domain and the dashed lines show the center of the object at different time slots.

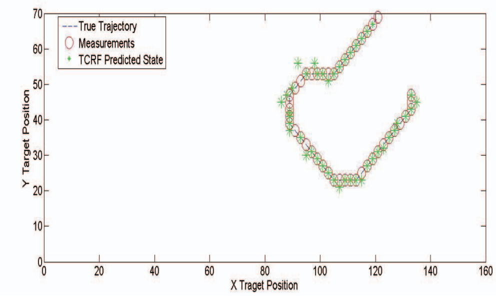

Figure 2: Simulated Motion: The moving object (black circle) has a trajectory over 35 frames, starting at the top-right of the image. The blue dashed line shows the true trajectory, the red circles the measurements, and the green stars the TCRF predicted state.

Our proposed method is evaluated by both real and simulated data. For the simulated motion, a black disk was moved on a white background, rendered into frames having a 120 × 160 resolution. The simulated motion dynamics are plotted in Figure 1, with the corresponding TCRF prediction in Figure 2. The results show that the TCRF prediction error is zero when the object velocity does not change; if the motion dynamics change we have prediction error at the time that the change occurs.

A separate comparison of the TCRF is shown in Table 1, where the prediction of the TCRF is compared with that of the Kalman filter. The strength of the TCRF becomes clear, in that a simple potential function is able to produce credible predictions, with an error much smaller than that of the Kalman filter. We examined quite a large number of feature functions. Our experiments show that not all selected features improve TCRF prediction. For this paper, only a combination of the reported features was utilized. The reader is reminded that since the DCRF does not posses any prediction step, we can not compare our proposed method with DCRF.

Finally, we examined our algorithm on real data, selected from standard datasets. The two selected sequences are shown in Figure 3 and Figure 4. In each evaluated sequence we show the prediction ability of the TCRF (top) followed by heuristic template-matching update step (bottom). In both experiments, it can be seen that the prediction of the TCRF and resulting heuristic update are very nearly equivalent, meaning that the TCRF alone accomplishes the bulk of the tracking

Table 1: MSE of prediction for Kalman filter and our TCRF method examined with two simulated motions.

Simulation No.	TCRF	Kalman Filter
Motion 1	0.8977	135.0659
Motion 2	0.6739	1.5021

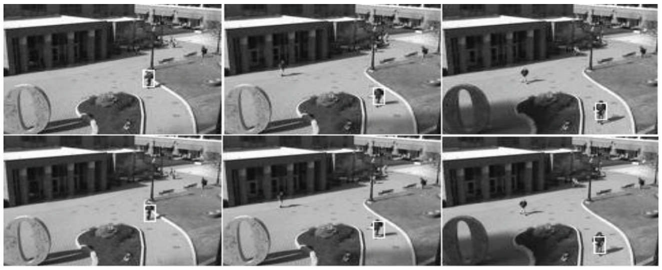

Figure 3: Real Sequence 1: the results from 3, 10 and 15 are shown. Top row: the TCRF prediction. Bottom row: further template matching, on the basis of the prediction in the top row. The prediction step alone accomplishes quite credible tracking, meaning that the template-matching update contributes relatively little to the tracking accuracy. (The examined dataset was obtained from www.cse.ohio-state.edu/otcbvs-bench)

task. It is worth noting the robustness of the TCRF, in the sense that the first dataset has background changes over time (cloud shadow) and object appearance changes in the second dataset.

6. CONCLUSIONS

In this paper we proposed a novel modification to CRFs to make them suitable for visual object tracking. The object motion is estimated using two consecutive frames (training phase) and the trained model is utilized to predict the position of the object in the following frames. The novelty of the our algorithm stems from the fact that it exploits temporal features, such as optical flow, in the CRF potential functions. This paper demonstrated the feasibility of temporal processing with CRFs, and specifically that the proposed TCRF is able to give credible tracking predictions, an important property has not yet been studied for the CRFs.

7. REFERENCES

[1] L. Taycher, G. Shakhnarovich, D. Demirdjian, and T. Darrell, "Conditional random people: Tracking humans with crfs and grid filters," *IEEE CVPR*, 2006.

[2] L. Zhang and Q. Ji, "Segmentation of video sequences using spatial-temporal conditional random fields," *Workshop on Motion and video Computing*, 2008.

[3] Y. Wang and Q. Ji., "A dynamic conditional random field model for object segmentation in image sequences," *IEEE CVPR*, 2005.

[4] L. Sigal, Y. Zhu, D. Comaniciu, and M. Black. , "Tracking complex objects using graphical object models," *Springer LNCS 3417*, 2004.

[5] V. Ablavsky, A. Thangali, and S. Sclaroff, "Layered graphical models for tracking partially-occluded object," *IEEE CVPR*, 2008.

[6] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," *IEEE ML*, 2001.

[7] C. Sutton, A. McCallum, and k. Rohanimanesh, "Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data," *IEEE MLR*, 2004.

[8] J. Yin, "Spatio-temporal event detection using dynamic conditional random fields," *IEEE AI*, 2009.

[9] J. Barron and N. Thacker, "Tutorial: Computing 2d and 3d optical flow," *ISBED., University of Manchester Stopford Building, Tina Memo No. 2004-012*, 2005.

[10] S. Geman and D. Geman, "Stochastic relaxiaon, gibbs distributions, and the bayesian restoration of images," *IEEE TPAMI*, 1984.

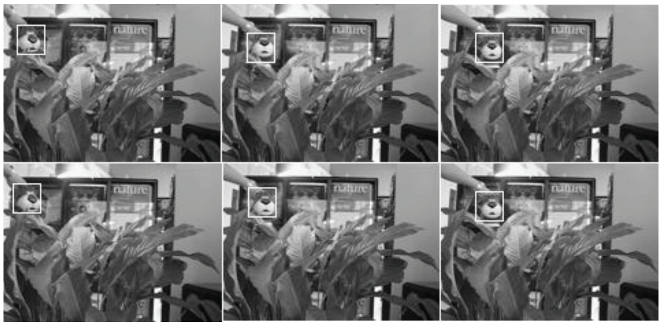

Figure 4: Result similar to Figure 3: Top row shows TCRF prediction in frames 3, 8 and 11. Bottom row shows template matching results. The object being tracked, the animal's face, is changing, since the face is turning. (This dataset was copied from www.vision.ucsd.edu/ bbabenko)