**Computer Communication & Collaboration (2013)**
**Submitted on 27/May/2013**

# How Conditional Random Fields Learn Dynamics: An Example-Based Study

***Mohammad Javad Shafiee***
School of Electrical & Computer Engineering, Shiraz University, Iran
E-mail: mj_shafiee@cse.shirazu.ac.ir

***Zohreh Azimifar*** (Correspondence author)
School of Electrical & Computer Engineering, Shiraz University, Iran
E-mail: azimifar@cse.shirazu.ac.ir

***Paul Fieguth***
Department of Systems Design Engineering, University of Waterloo, Canada
E-mail: pfieguth@uwaterloo.ca

**Abstract:** In this paper[1]we investigate how Conditional Random Fields (CRFs) learn dynamics. To demonstrate the ability of CRF in learning dynamics, a discriminative probabilistic framework, Temporal Conditional Random Fields, is presented for the modeling of the object motion and tracking. The main drawback of generative models, such as HMM and MRF is that they can simply employ the relation between states without considering the relation between states and measurements, while discriminative frameworks can model any arbitrary relation between measurements and states. To facilitate such a powerful graphical model to learn the object motion, and to achieve a CRF-based estimation based upon the advantage of the discriminative framework, a set of graphical temporal relations is proposed for the object tracking, including feature functions, such as optical flow (calculated based upon consequent frames) and line field features. Based on temporal feature function, we show the ability of CRF in dynamic learning mathematically. As it is assumed that the object motion is nearly constant and that the current measurement is not available when TCRF estimates the target state, therefore, the changing of the object motion is addressed by utilizing a template matching in order to determine and then retrain TCRF. The proposed method is validated using synthetic and real data sequences. This shows that the TCRF estimation error is approximately zero.

**Keywords:** Visual Tracking, Motion Dynamic, Discriminative Models, Conditional Random Fields, Potential Function

## 1 Introduction

Event modeling has attracted a large body of research during the past two decades. Due to the measurements that are corrupted or insufficient, the real event modeling is an ill-posed problem. To address this drawback some types of prior knowledge, regularization or constraint can be applied to make the problem tractable. Object tracking, image de-noising and surface reconstruction are the instances of the ill-posed problems.

[1]The primary idea of this work was published at [1]

Our objective in this paper is to model the target dynamic for the purpose of object tracking. A significant number of heuristic and statistical methods have been proposed to solve this problem. The most common method is known as Kalman filtering. Kalman filter performs an optimum least-squares estimate in the presence of the Gaussian measurement noise. The Kalman filter and its variations predict the next object state by means of a predefined dynamic, followed by updating the predicted state using the new measurement. Within the context of the graphical modeling, the Kalman filter behavior is identical to that of Hidden Markov Models (HHM). This model is a generative approach modeling the joint distribution of measurements and labels. Markov Random Fields (MRFs) are also generative models, assuming the conditional independence between measurements when conditioned on labels (states).

Generative approaches such as MRF, HMM and especially Kalman filter utilize a prior model i.e. state probability density. Therefore, they simply model the relation between states without considering the relation between the states and measurements. In other words, they estimate the joint probability of measurements and states based on the two probabilities of the prior model and likelihood model in which the relation of states between measurements are not considered.

From a broader viewpoint, there exist various statistical approaches addressing such problems. According to the most well-known categorization in the context of Bayesian graphical modeling, these models are divided into generative and discriminative ones, depending upon their relaxation assumptions. The assumptions, characteristics, and computational complexity of each model lead to different applications for each framework.

Generative models estimate the joint probability (relationship) of measurements and states. According to the Bayes theorem, this joint probability is equivalent to the product of the prior states probability and conditional probability of measurements, given the states. In other words, the modeling with generative structures requires the prior model, e.g. the Kalman filter needs a predefined dynamic. Modeling all dependencies using the conditional probability distribution of measurements given the state is, typically, highly complex; thus, all measurements are assumed to be conditionally independent, given the states. The conditional independence assumption relieves the heavy computational burden, considering the cost of reducing the accuracy of the modeling problem where the state measurement interactions cannot explicitly be ignored.

Discriminative models such as Conditional Random Field (CRF), on the other hand, relax the independence assumption of the generative methods by directly modeling the conditional probability distribution of states given measurements. In this approach, it is not required that the prior model be exploited explicitly; that is, the discriminative model spots the prior model implicitly. The discriminative model tends to employ a log-linear model based on a number of energy functions which are weighed using some real values.

Many image processing problems limit data for the purpose of generative modeling, leading to comparatively less accurate MRF models. On the other hand, the CRFs can solve a wider range of computer vision problems due to their explicitly modeling the conditional distribution more efficiently without any explicit requirement for the prior model.

Considering the above-mentioned fact, our objective is how CRF can learn different dynamics. To illustrate the ability of CRF a new probabilistic approach to the object tracking is proposed based on CRF. The main objective of the proposed framework is the modeling of the interrelationship between past/current states and past measurements to learn the object dynamic and, therefore, to estimate the object state in the application of tracking. In a discriminative framework, a small number of frames are used to model the object motion by the Temporal Conditional Random Field (TCRF) framework. Following the learning of the TCRF state and the finding of weights corresponding to each feature function, the TCRF is used to estimate the next target state. Experiments indicate that TCRF can efficiently learn the target dynamic on-line and estimate the target state without any error should the target dynamic remain unchanged.

This paper is organized as follows: Section 2 reviews the related literature. In Section 3 CRFs

are described and the TCRF method is proposed. Section 3 also shows how CRF can learn object dynamics mathematically. Section 4 explains the proposed framework a step at a time. Afterwards the experiments are discussed and the conclusion is presented.

## 2 Related Work

In this paper, TCRF, which is a target state estimator based on a statistical model of previous measurements and the previous and current state, is proposed. To the best of our knowledge, no research has been conducted in the field of the CRF visual tracking when the current measurement is not available. However, CRF has been utilized as an estimation tool in other fields. Taycher*et al.* [2] proposed human tracking based on a CRF, with a$L_1$ similarity measure as the potential function. In [2] different poses are considered as states for tracking within a sequence of images, where the number of states has previously been determined. CRFs were also applied to image-sequence segmentation [3, 4], where the random fields are modeled using spatial and temporal dependencies. Sigal*et al.* [5] use the two-layer spatio-temporal models for the component-based detection and tracking of objects in video sequences. Each object or component of an object is considered as anode of a graphical model at a given time and the graph edges represent the learned spatial and temporal constraints. Considering this, Ablavsky*et al.* [6] proposed a layered graphical model for the partially occluded object tracking. A layered image-plane represents motion around a known object that is associated with a pre-computed graphical model. Ren[7] proposed a framework to find all people in the archived films. The proposed framework finds people in low-quality images, motion blur, partial occlusion, non-standard poses and crowded scenes. His tracker performs one dimensional CRF (chain) to integrate information across frames and to re-score the tentative detections in trajectories.

In [8] target silhouette is tracked on a video sequence. The proposed algorithm fuses different visual cues by using a conditional random field. The temporal color similarity, the spatial color continuity and the spatial motion continuity are the CRF feature functions employed in this work.

Our objective here is to model the object motion in a simple yet general manner. It is assumed that the video frame rate is high; therefore, there is no abrupt change in the position and motion direction of the object along the consecutive frames. Initially the motion conditional distribution $P(Y|M)$is modeled by TCRFusing the previous frames, where the measurements are shown as $M$and estimated states as$Y$. Then the object position in the current frame can be estimated given the previous frames. TCRF is adapted to estimate the object state at time $t$when the measurement of $t$is not yet available. It was empirically observed that the estimation tends to have no error when no change occurs in the object motion.

To address the object motion changing issue, following the state estimation and in case of the current availability, a heuristic procedure searches around the estimated coordinates to find the best matching sub-image with an extracted target template using the two last training frames. Should the TCRF estimation for time $t$have significant difference, as compared with the template matching coordinates, it is assumed that the changed target motion and the TCRF should be retrained with frames $t-1$and $t$.

## 3 Temporal Conditional Random Fields

This paper investigates the TCRFs with feature functions describing temporal relations between successive frames. Alternatively, our objective is to investigate how to use the CRF with dynamic motion learning. It must be pointed out that the object tracking has a vast literature, spanning many years. The purpose of this paper is not to contest that literature; rather, this paper aims a studying the effectiveness and potentials of the CRF methods in tracking, specifically in motion dynamic learning. Our goal is to study several can did a potential functions, and to assess their ability in the context of CRF tracking. This research indicates how the feature functions simply extracted from the field contribute to an efficient tracking by CRFs.

### 3.1 Conditional Random Fields

The idea of a conditional random field was first proposed by Lafferty *et al.* [9]. 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 probability distribution in this Bayesian modeling approach is obtained based on the Principle of Maximum Entropy [10]. The basic idea of Maximum Entropy Models is to find the conditional probability of states given measurements by taking into account the largest possible conditional entropy states given measurements of the maximum consistency with the information extracted from training materials. This idea leads to a log-linear model with feature functions and weights corresponding to each function. Motivated by Maximum Entropy Models, CRF is introduced to directly model the conditional probability distribution of labels given measurements without considering the conditional independence assumption.

Much like Markov Random Fields (MRF), CRF is an undirected graphical model with each node representing a random variable, and each edge measuring the relation and dependency between random variables. CRF is technically defined as an undirected graph $G(V,E)$wherein each vertex $v \in V$corresponds to each of the random variables representing a$Y_v \in Y$and each $e \in E$shows the dependency between two end point nodes (random variables) [11]. The set $\{X,Y\}$is a CRF if each random variable $Y_v$, when conditioned on $x \in X$obeys the Markov property with respect to graph $G(V,E)$where $X$is the observation and $Y$is the label (target) variable.

The general model formulation of CRFs is:

$$P(Y|X) = \frac{1}{Z(X)} \prod_{c \in C} \psi_c(Y_c, X_c) \qquad (1)$$

Where $\psi_c(X_c, Y_c)$is a potential function corresponding to cliques $c \in C$[2]. Clique templates make assumptions on the structure of the underlying data by defining the composition of the cliques.}, $Z(X)$ is normalization constant -- partition function -- with respect to all the possible values of target variable$Y$:

$$Z(X) = \sum_{Y'} \prod_{c \in C} \psi_c(Y'_c, X_c) \qquad (2)$$

The potential function $\psi_c$is an arbitrary non-negative function of $X_c$ and $Y_c$. According to the Maximum Entropy Model[3], potential functions are formulated as exponential functions of weighted features. Such a formulation is commonly used because it satisfies the strict positivity of the potential functions [10]. The log-linear model of (1) is formulated as:

$$P(Y|X) = \frac{1}{Z(x)} \prod_{c \in C} \exp\left(\sum_{k=1}^{K(c)} \lambda_k(c) f_k(Y_c, X, c)\right) \qquad (3)$$

$$Z(X) = \sum_{Y'} \prod_{c \in C} \exp\left(\sum_{k=1}^{K(c)} \lambda_k(c) f_k(Y'_c, X, c)\right)$$

where$f_k(Y_c, X_c, c)$ and $\lambda_k(c)$ represent the $k^{th}$ real-valued feature function defined on the clique $c$ and its corresponding model parameter, respectively. The number of feature functions defined over clique $c$ is determined by $K(c)$, which has an arbitrary value for each clique $c$.

After defining the CRFs and considering the fact that it has been applied successfully in

[2] A clique $c$ is a set of nodes $Y_c$ in$G$, such that each pair $(v_i, v_j) v_i \in c$ and $v_j \in c$ are connected by an edge in $G$. It would be mentioning that a single node is also considered a clique [12].

[3] With incomplete information about a probability distribution, the best unbiased distribution is as uniform as possible by the available information. Then the best distribution is the one which maximizes the entropy given the constraints from the training material [10].

various machine learning problems the TCRF, which utilizes 2D CRF to model the spatio-temporal relations between successive frames, is proposed. TCRF models the relation between successive frames according to a defined neighborhood in the spatial and temporal domain. In this paper, TCRF simply employs two consequent frames for training according to the first-order Markov assumption. By increasing the number of frames TCRF can learn different dynamic motions.

### 3.2 CRF Evolution in Time

Following the early CRFs, which consider simply the spatial relations between the random fields, Sutton *et al.*[13] proposed dynamic conditional random fields (DCRF) to capture the spatial relations between the neighboring nodes and the temporal relations across temporally separated frames [14]. One can interpret [15] DCRF as a kind of CRF that has repetitive structures and parameters over time.

This study examines the temporal conditional random fields associated with the spatial and temporal feature functions which represent the relationship between nodes in the time domain as well as the spatial domain. Our objective is to model the temporal relation between the two successive frames in order to add the motion dynamic learning to the CRF framework. In other words, this research deals with the CRF ability of to estimate the target state without the corresponding current measurement in visual tracking. One frame is segmented using the trained TCRF model, as well as the previous state and frame as measurements. The relation between node states in temporal domain is modeled by incorporating some feature functions based on the labeled frame at time $t$and $t+1$. This means that the proposed framework can segment the frame at time $t+1$into the target and background merely by using frame at time $t$.

The goal of tracking is to estimate $Y_\tau$based on $Y_t$ $(t<\tau)$and $m_{t'}$ $(t'\leq\tau)$. Therefore, the state $Y_{t+1}$ can be estimated by the CRF conditioned on $Y_{1:t}$and $m_{1:t+1}$:

$$P(Y_{t+1}|Y_{1:t},m_{1:t+1})=\frac{1}{Z(m_{1:t+1})}exp\left(\sum_{y_{t+1,i}\in Y_{t+1}}\left\{\sum_k^K\lambda_kF_k(y_{t+1,i},Y_{1:t},m_{1:t+1},N_i)\right\}\right) \quad (4)$$

where $y_{t+1,i}$ shows the binary label (foreground or background) of any pixel $i$ within the frame $t+1$. $Y_{1:t}$are the binary labels fields from the initial time to the time $t$, and $m_{1:t+1}$ are the observations from the initial time to the time $t+1$. Feature function$F(y_{t+1,i},Y_t,G(m_{t+1}),N_i)$ is an arbitrary function, where $N_i$is a set of neighbors for each node $i$. Figure 1 shows Eq. (4) graphically.

According to the first-order Markov property, Eq. (4) can be rewritten as:

$$P(Y_{t+1}|Y_t,m_{t+1})=\frac{1}{Z(m_{t+1})}exp\left(\sum_{y_{t+1,i}\in Y_{t+1}}\left\{\sum_k^K\lambda_kF_k(y_{t+1,i},Y_t,m_{t+1},N_i)\right\}\right) \quad (5)$$

The advantage of CRF is to select any arbitrary feature function based on application. Based on this virtue of CRF, any combination function of the measurements and labels can be exploited.

The feature function can describe any arbitrary relation between measurements and states (labels). To generalize this statement, the feature function can describe any functionality between states and measurements. In other words, instead of using the measurements themselves, any desired function of the measurements can be utilized. This is an advantage of CRF, in which:

$\sum_k^K\lambda_kF_k(y_{t+1,i},Y_t,G(m_{t+1}),N_i)$ is the general form of $F_k\left(y_{t+1,i},Y_t,m_{t+1},N_i\right)$.

A function of measurement $G(m)$can be used when the primitive measurement $m$ is not directly useful. Accordingly, the (5) is rewritten as follows:

$$P(Y_{t+1}|Y_t,m_{t+1})=\frac{1}{Z(m_{t+1})}exp\left(\sum_{y_{t+1,i}\in Y_{t+1}}\left\{\sum_k^K\lambda_kF_k(y_{t+1,i},Y_t,G(m_{t+1}),N_i)\right\}\right) \quad (6)$$

Our objective is to learn the target motion dynamic or to estimate the target state at time $t+1$ in the absence of the measurement at time $t+1$ simply by using the measurement at time $t$. As mentioned above, CRF can utilize any function of the measurements. Also, when it is assumed that video frame rate is high and when there is no sudden dynamic change, the following approximation, eq.(7), can be obtained:

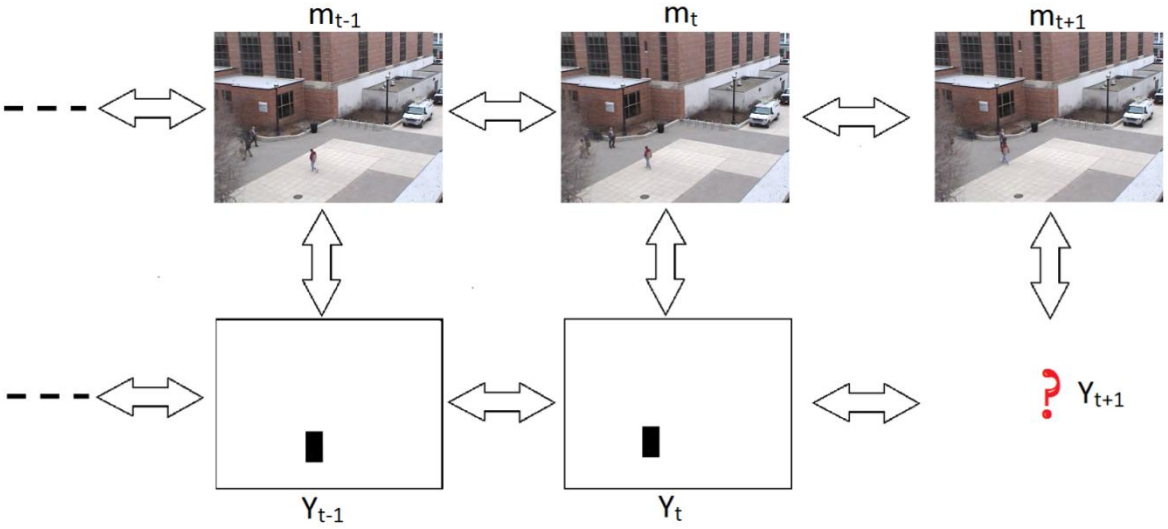



**Figure 1** In the general form, the tracking goal is to estimate the state $Y_{t+1}$based on the previous states at time $1{:}\,t$and the observed measurements $(1{:}\,t+1)$.

$$m_t \cong g(m_{t+1}) \qquad (7)$$

where function $g(.)$ belongs to the general family of $G(.)$ and generates $m_t$ from $m_{t+1}$. Now $g(m_{t+1})$ can be utilized as one of $G(.)$:

$$P(Y_{t+1}|Y_t, m_{t+1}) = \frac{1}{Z(m_{t+1})} exp\left(\sum_{y_{t+1,i\in Y_{t+1}}} \left\{\sum_k^K \lambda_k F_k(y_{t+1,i}, Y_t, g(m_{t+1}), N_i)\right\}\right)$$

$$Z(m_{t+1}) = \sum_{Y_{t+1}\in Y'} \left(\sum_{y_{t+1,i\in Y_{t+1}}} \left\{\sum_k^K \lambda_k F_k(y_{t+1,i}, Y_t, g(m_{t+1}), N_i)\right\}\right) \qquad (8)$$

where $Y'$is all configurations of $Y_{t+1}$. According to (7), $g(m_{t+1})$ is replaced by $m_t$:

$$P(Y_{t+1}|Y_t, m_{t+1}) \approx \frac{1}{Z(m_{t+1})} exp\left(\sum_{y_{t+1,i\in Y_{t+1}}} \left\{\sum_k^K \lambda_k F_k(y_{t+1,i}, Y_t, m_t, N_i)\right\}\right)$$

$$Z(m_{t+1}) \approx \sum_{Y_{t+1}\in Y'} \left(\sum_{y_{t+1,i\in Y_{t+1}}} \left\{\sum_k^K \lambda_k F_k(y_{t+1,i}, Y_t, m_t, N_i)\right\}\right) \qquad (9)$$

The right side of (8) is equal to $P(Y_{t+1}|Y_t, m_t)$, therefore, it can be said that $P(Y_{t+1}|Y_t, m_{t+1})$ is approximated by $P(Y_{t+1}|Y_t, m_t)$. In the proposed TCRF, one frame is considered as a measurement used to estimate the next target state as shown in Figure 2. Since it has been assumed the frame rate videos are high, a simple relation between frames is sufficient to model an object motion. Therefore, based on (8)$P(Y_{t+1}|Y_t, m_{t+1})$ can be approximated by using the proposed TCRF:

$$P(Y_{t+1}|Y_t, m_{t+1}) \approx P(Y_{t+1}|Y_t, m_t) =$$
$$\frac{1}{Z(m_t)} exp\left(\sum_{y_{t\in Y_{t+1}}} \left\{\sum_{k_1=1}^K \lambda_{k_1} f_{k_1}(y_{t+1,i}, m_t, N_i) + \sum_{k_2=1}^{K'} \lambda_{k_2} f_{k_2}(y_{t+1,i}, Y_t, N_i)\right\}\right) \qquad (10)$$

Our TCRF employs two kinds of feature function $f(y_{t+1,i}, m_t, N_i)$ and $f(y_{t+1,i}, Y_t, N_i)$ as a single potential function and a temporal interaction potential function, respectively. Here $N_i$ is a set of temporal neighbors for each node $i$. The interaction potential function utilizes pairwise cliques; thus, the TCRF definition can be generalized by using the neighbors of each node rather than a clique. The neighbors for each node are a combination of several cliques.

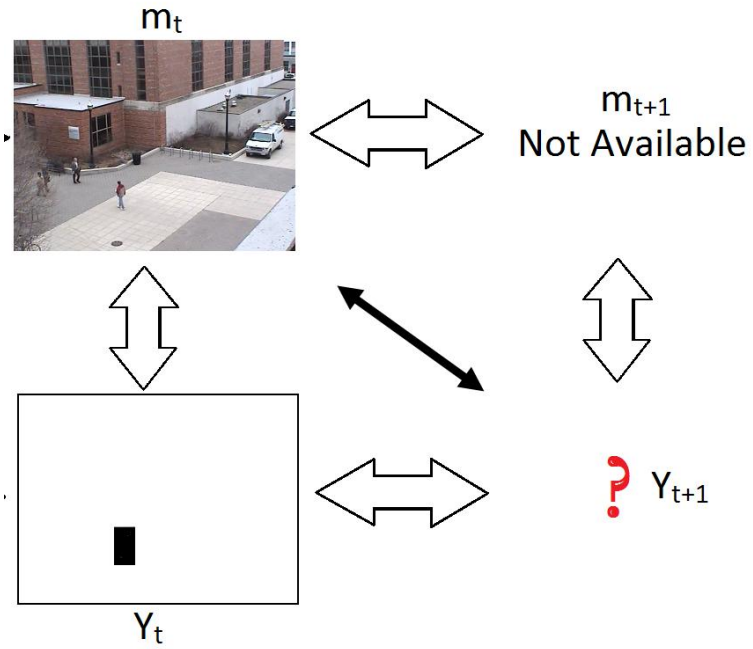



**Figure 2** TCRF estimates the state at time $t+1$ based on the measurement and the state at time $t$

The goal is to study the effects of the different potential functions in modeling an object motion in the context of CRFs. As stated before, any arbitrary feature function can be applied to model the conditional probability of states given measurements. It is crucial that feature functions with maximum discriminative property be utilized. As mentioned above, any desired function of measurement can be employed when the primitive measurement is not available such as the motivation of this study. Since tracking problems are inherently temporal, the potential functions are required to be derived based on the motion features with some temporal dependency. The rest of this section explains the functions to be used in our TCRF framework.

## 3.3 Feature Functions

In this research we utilize different feature functions to model the object motion by CRF. Our studies show that to learn the target dynamic motion with TCRF, we need feature functions that indicate both the object motion direction and the object displacement with most discrimination. We examine our system with a number of feature functions and finally select the most appropriate ones. Our selection criteria are simplicity and effective of feature functions.

### 3.3.1. Optical Flow

The single potential function is described with two values showing the velocity of each pixel in both $x$ and $y$ directions in two adjacent frames which are estimated by optical flow [16].

Optical flow is an approximation of motion based upon local derivatives in a given the sequence of images [16]. This approximation specifies how far each pixel moves in two adjacent images. First assumption by optical flow is that any pixel intensity from one frame to the next obeys this property:

$$I(x,y,t) = I\big((x+\delta x),(y+\delta y),(t+\delta t)\big) \qquad (11)$$

where$I(x,y,t)$is the pixel intensity at position of $(x,y)$and at time $t$ and $\{\delta_x,\delta_y,\delta_t\}$ indicates 2D velocity in plane. When we apply first order Taylor, velocities in two directions are obtained as:

$$I(x,y,t) = I(x,y,t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\,\delta t \qquad (12)$$

Substituting (12) in (11) we obtain:

$$\frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\,\delta t = 0 \qquad (13)$$

We rewrite above formula to:

$$\frac{\partial I}{\partial x}\frac{\delta x}{\delta t} + \frac{\partial I}{\partial y}\frac{\delta y}{\delta t} + \frac{\partial I}{\partial t}\frac{\delta t}{\delta t} = 0 \qquad (14)$$

where$\delta_x/\delta_t$ and $\delta_y/\delta_t$ are pixel velocity in directions $x$ and $y$, respectively. We consider these two values as single potential functions for each node of TCRF (pixel). By rewriting (14) one obtains:

$$(I_x, I_y).\big(v_x, v_y\big) = 0 \qquad (15)$$

Where $(I_x, I_y)$is the spatial intensity gradient.According to the section 4.1, optical flow is computed based on $Y_t$ (the current target state) and synthetic images (the next target state) which is created in MAP estimation. In other words, optical flow is a feature function which describes the temporal relation of between label nodes (state) of $Y_t$ and $Y_{t+1}$.

### 3.3.2. Line Field

Line fields were first introduced by S. Geman and D. Geman [17], 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 given in [17]:

$$F(Y_t, Y_{t+1}, i, N_i) = \sum_{j\in N_i, y\in Y} 1 - \delta(y_t(i) - y_{t+1}(j)) \qquad (16)$$

where$\delta(a)$ is Kronecker delta function. We also exploit the duality of feature functions in order to reduce the similarity between the function value of each configuration in temporal relation neighbors and to reinforce feature functions be more discriminative. Duality is graphically depicted in Figure 3. The dual form of (16) is:

$$F(Y_t, Y_{t+1}, i, \overline{N}_i) = \sum_{j\in \overline{N}_i, y\in Y} 1 - \delta\big(y_{t+1}(i) - y_t(j)\big) \qquad (17)$$

where$N_i$and $\overline{N_i}$are the neighborhood sets of $i$in fields$Y_{t+1}$ and $Y_t$, respectively.

### 3.3.3. Ising

The Ising model [17] 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 the potential form as

$$F\big(m_t, Y_{t+1,N_i}, i\big) = \sum_{j\in N_i, y\in Y} m_t(i) \times y_{t+1}(j) \qquad (18)$$

with the following dual form:

$$F\big(m_t, Y_{t+1,\bar{N}_i}, i\big) = \sum_{j\in \bar{N}_i, y\in Y} m_t(j) \times y_{t+1}(i) \qquad (19)$$

Since label $Y_{t+1}$is binary $\{-1,1\}$ and the object appearance is nearly equivalent in the consequent frames Ising function has negative value in object region and positive in background region.

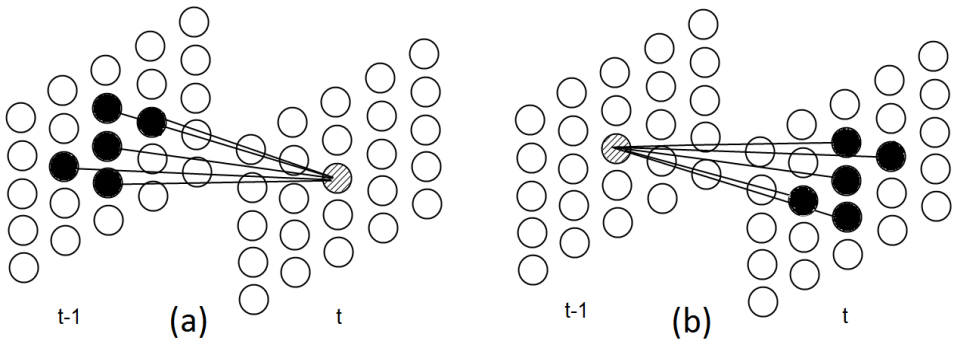



**Figure 3** Duality relation in the feature functions. (a) neighbors of gray circle at time $t$ is in time $t-1$black circle). (b) shows neighbors for the gray circle at time $t-1$in time $t$. This forms of the neighborhood relation is used for the dual feature function.

### 3.4 Training

Maximum likelihood is a common method to estimate the parameters of CRFs. Training is done by maximizing log-likelihood $l$ on the training data:

$$l(\lambda) = \sum_{y_{t\in Y_{t+1}}}\{\sum_{k_1}^{K}\lambda_{k_1}f_{k_1}(y_{t+1,i}, m_t, N_i) + \sum_{k_2}^{K'}\lambda_{k_2}f_{k_2}(y_{t+1,i}, Y_t, N_i)\} - \log(Z(m_t)) \quad (20)$$

Because the log-likelihood function $l(\lambda)$ is concave, the parameters $\lambda$ can be chosen such that the global maximum is obtained and the gradient or vector of partial derivatives with respect to each parameter $\lambda_k$ becomes zero. Differentiating $l(\lambda)$ with respect to parameter $\lambda_k$ gives:

$$\frac{\partial l}{\partial \lambda_{k_1}} = \sum_{y\in Y}\left(f_{k_1}(y_{t+1,i}, m_t, N_i) - \sum_{Y\prime} f_{k_1}(y_{t+1,i}, m_t, N_i)P(Y'|X)\right)$$

$$\frac{\partial l}{\partial \lambda_{k_2}} = \sum_{y\in Y}\left(f_{k_2}(y_{t+1,i}, Y_t, N_i) - \sum_{Y\prime} f_{k_2}(y_{t+1,i}, Y_t, N_i)P(Y'|X)\right) \quad (21)$$

An exact solution does not exist, therefore, the parameters are determined iteratively using gradient descent. Our TCRF training is performed by Belief Propagation method [18]. The training of TCRF is not our concern in this paper. The TCRF is trained initially by two frames at time $t=0$ and $t=1$. Figure 5 shows the TCRF training schematically.

### 3.5 Inference and Decoding

After training the TCRF, given the observation $m_t$, evaluating probability of each random variable in represented graph is called inference and the task of assigning the output variable $Y$ -determining states with maximum probability- is decoding. Eq. (22) and (23) show formal definition of inference and decoding, respectively:

$$P_{y_i} = P(Y_{t+1} = y_i | m_t, Y_t) \quad \forall y_i \in Y \quad (22)$$

$$Y^* = argmax_Y P(Y_{t+1} | m_t, Y_t) \quad (23)$$

Because our proposed TCRF utilizes feature functions as well as their dual form, inference imposes a large computational burden. To overcome this problem, we assume that object motion does not have sudden change and TCRF considers a combination of inference and decoding with some modification that is explained next.

## 4 Tracking with TCRF

Given a temporal potential function, the temporal CRF can be created to estimate object position in the future frames. After estimating the object position and true measurement arrives, a heuristic method (such as the template matching) searches near the estimated position to find the coordinates of the best matched candidate. The template matching is performed to address the changing of the object dynamic issue and determines the change of the object motion in case of the availability of the true measurement. Figures 5, 6 show the two stages of training and estimation by TCRF schematically. Should the estimated and matched coordinates be very different, it is assumed that the object motion changes. As a result the TCRF training is repeated using the two last frames. The diagram depicted in Figure 4 shows the proposed tracking model.

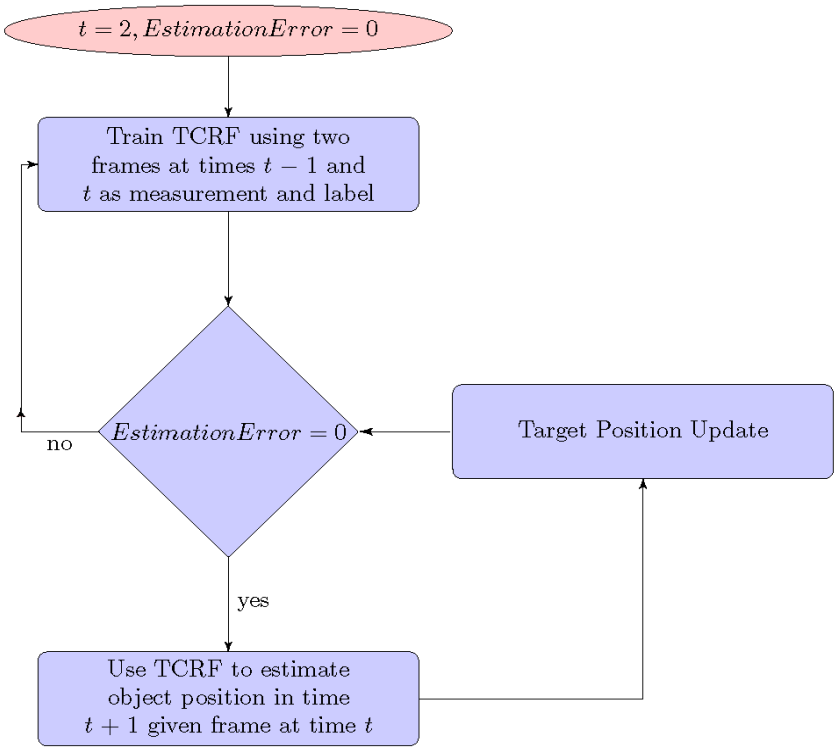



**Figure 4** Flowchart describes the proposed tracking system: should the estimation error be non-zero, TCRF is trained with the two last frames.

### 4.1 MAP Estimation

After training the TCRF, we use maximum a posteriori (MAP) to estimate the object position at time $t+1$. Since the TCRF models the object motion, the estimation is performed by evaluating the probability of the next target location around its previous position at time $t$. A set of synthetic images with a synthesized target are created. The synthesized target obeys a variety of dynamics, therefore, the probability of each image assessed by the TCRF. The estimated target position (in the frame $t+1$) is found from that sample with the maximum probability. That is, the synthetic image that maximizes the TCRF probability shows the segmentation of the next frame as foreground and background.

### 4.2 Determination of change on object motion

To overcome the object motion changing problem, a template matching procedure is utilized to specify a change in times of occurrence. Should the TCRF target estimated state be different from template matching estimated coordinate, this indicates that the object motion has changed and that TCRF must be retrained. It would be worth mentioning that template matching procedure is applied at a center that is specified by the TCRF estimated coordinate in the frame $t+1$ searching and obtains the object position. TCRF simply uses the frame at time $t$ as the measurement. In contrast, the template matching employs the frame $t+1$ to estimate object target at time $t+1$.

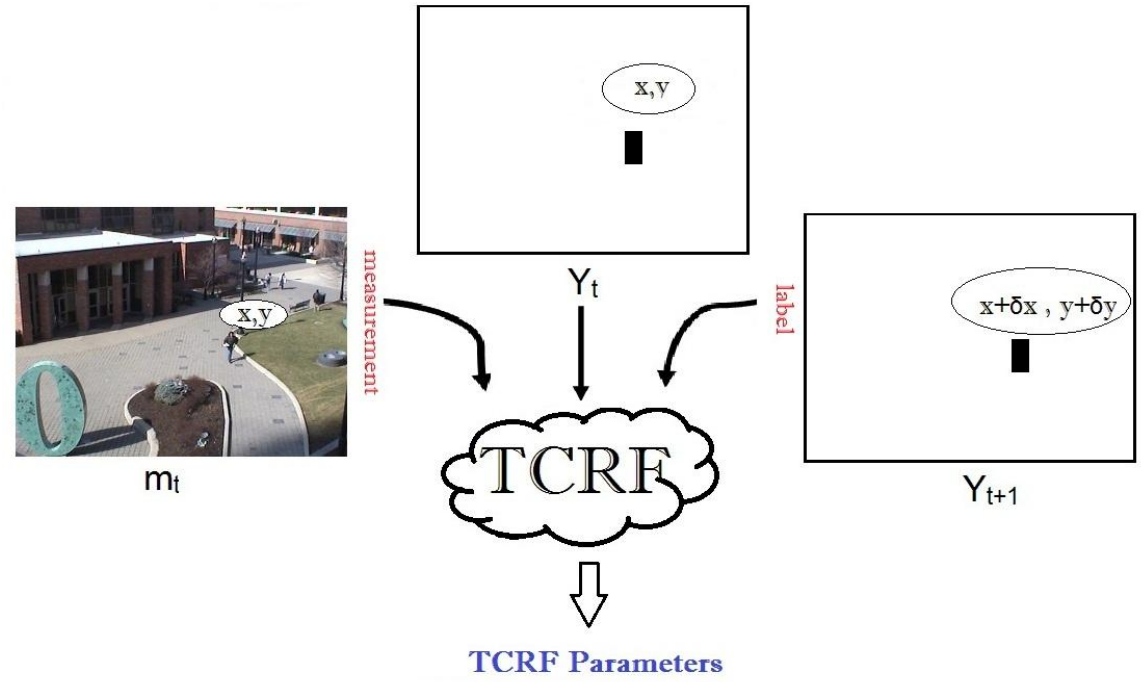



**Figure 5** To train TCRF, the frame at time $t+1$ is segmented into the foreground (black) and the background (white) to create the label $Y_{t+1}$. The frame $t$ ($m_t$) and the segmented frame $t+1$ ($Y_{t+1}$) are fed into the TCRF for training.

## 5 Results and Discussions

TCRF is initially trained with the first two frames. After that, TCRF can estimate the next target state at time $t+1$ based on the current measurement and state. In time $t+1$, a template matching is performed to evaluate the estimation result of the TCRF and determines the object motion change in case of occurrence. The TCRF estimation error is nearly zero until the motion dynamic changes. TCRF is always retained using the two latest frames when the motion dynamic changes (i.e. update TCRF).

In section 3.2 we show that how CRF can learn dynamic and therefore, predict next target state without the corresponding measurement. In this section we show this ability of CRF based on some examples. The proposed method is evaluated by both real and simulated data. To simulate the motion, a black disk was moved on a white background and rendered into the frames of $120 \times 160$ resolution. One of the simulated motion dynamics is plotted in Figure 7(a), with the corresponding

TCRF estimation depicted in Figure 7(a). The results show that the TCRF estimation error is zero when the object velocity does not change; should the motion dynamic change, the estimation error will increase dramatically when the change occurs. Figure8 also shows the estimation result for another maneuver target motion. Obviously, the TCRF estimation error is zero except that when the object dynamic is changed as shown in Figure 8(a). Figure 8(b) depicts the Kalman filter estimation result.

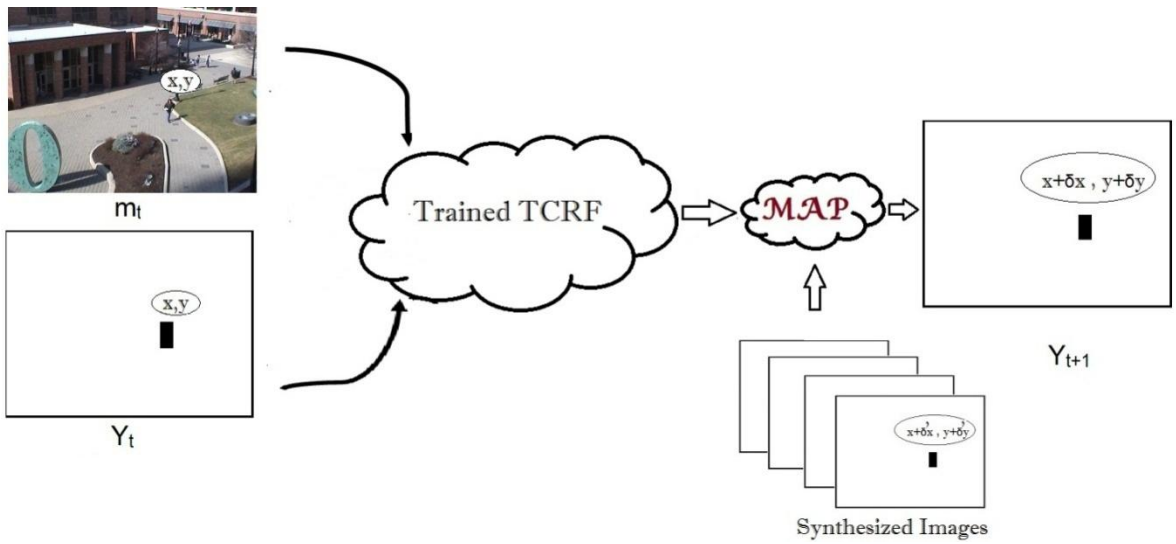
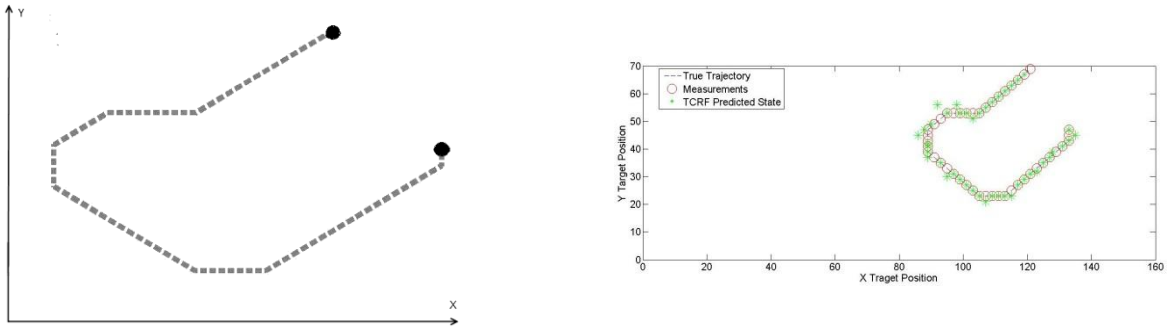


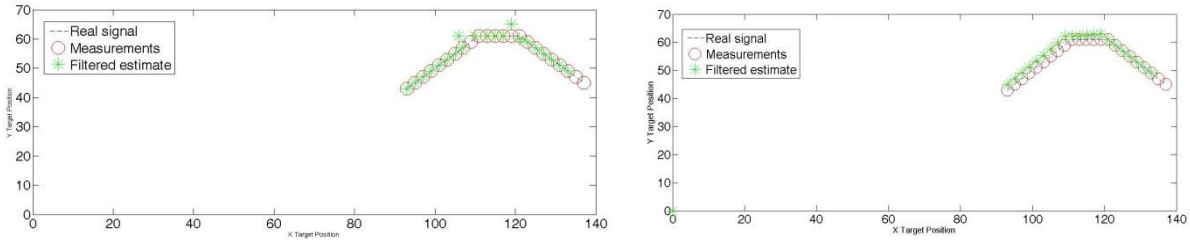

**Figure 6** To estimate the next target state ($Y_{t+1}$), the frame at time $t$ ($m_t$), the label at this time and the synthesized images are used to estimate $Y_{t+1}$. Section 4.1 describes how the proposed framework estimates the next target state.




(a) A sample of a maneuvering target used to evaluate the proposed method. The object motion starts at the top right of the domain where the dashed lines show the center of the object at the different time slots.

(b) 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 are measurements, and the green stars shows the TCRF estimated state.

**Figure 7** Simulated Motion: a maneuver target motion (a) and the TCRF estimation result (b).




(a) TCRF estimation result (b) Kalman estimation result

**Figure 8** Another sample of a maneuvering target used to test the proposed method. The object motion starts at right of the domain. (a) shows the TCRF estimation and (b) illustrates the Kalman filter result.

A quantitative evaluation of the TCRF is shown in Table 1, where the estimation of the TCRF is compared with that of the Kalman filter. The strength of the TCRF becomes clear in regard to the fact that a simple potential function is able to produce credible estimations with an error much smaller than that of the Kalman filter. It is noteworthy that the Kalman filter input is the true position of the object, and that in our experiments we assumed no noise on the measurements of the Kalman filter, while the proposed TCRF input is intensity value of a complete a video sequence frame. Despite other predictors, TCRF considers and processes the noise implicitly and automatically. This makes this algorithm highly effective, where the estimation error is very important and the algorithm computational complexity is not vital.

We examined quite a large number of feature functions. Our experiments indicate that not all selected features improve TCRF estimation. For this paper, only a combination of the reported features was utilized.

Finally, we evaluated our algorithm performance on real data, selected from standard datasets. The three selected sequences are shown inFigure9-11. In each evaluated sequence the estimation ability of the TCRF in the absence of current measurement (top) is shown followed by heuristic template-matching when the true measurement is ready(bottom).

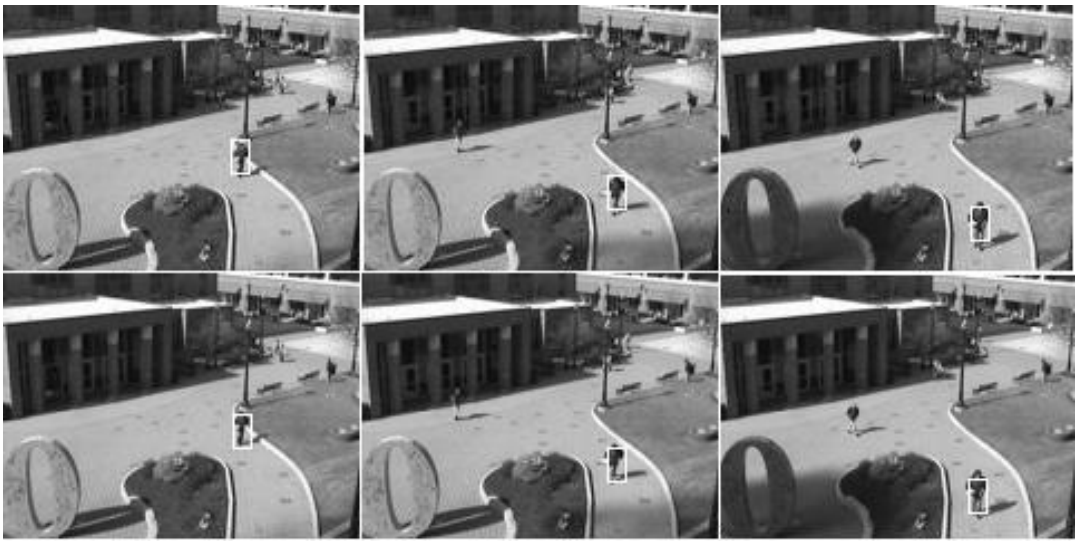
As shown in Figure 9 background is changed along the time and intensity of the pixels changes in each frame with shadow. Therefore, our proposed TCRF can estimate the next object position with background varies. Because the TCRF learned the object dynamic, the presence of other objects do not effect on the estimation result. Our reason to chosen Figure 10 as benchmark of our algorithm is variation of face appearance. The animal face turns along frames, therefore, face appearance changes along time. The last experiment was done on a moving human. In Figure 11 frame 6, man is in occlusion condition. As before, proposed TCRF works properly in this situation.

In these experiments, it can be seen that the estimation of the TCRF and the resulting of the heuristic method are very nearly equivalent, meaning that the TCRF alone accomplishes the bulk of the tracking 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 and occlusion scene in third dataset.

## 6 Conclusion

In this paper we proposed a novel modification to CRFs to make them suitable for the visual object tracking. The main objective of the paper is to illustrate how CRF can learn target dynamic that it is shown by motion estimation problem. The object motion is estimated using two consecutive frames (training phase) and the trained model is utilized to estimate the position of the

object in the following frames. The novelty of our algorithm stems from the fact that it exploits the temporal features (e.g. 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 in the absence of the current measurement, an important property has not yet been studied for the CRFs. As stated before, TCRF assumes the object motion has constant velocity between two frames. Our primary experiments show TCRF also can potentially be used in non-rigid object and multiple object tracking.



Real Sequence 1: the results from 3, 10 and 15 are shown.Top row: the TCRF estimation by the previous measurement and in the absence of current measurement. Bottom row: further template matching,on the basis of the estimation in the top row. The estimation stepalone 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).

**Table 1** MSE of the estimation for Kalman filter and our TCRF method examined with the different simulated motions.

| Simulation No. | TCRF | Kalman Filter |
|---|---|---|
| **Motion 1** | 0.8977 | 135.0659 |
| **Motion 2** | 0.6739 | 1.5021 |
| **Motion 3** | 0.6964 | 4.1154 |
| **Motion 4** | 0.2759 | 1.3258 |

**Figure 9** Result similar to Figure9: Top row shows TCRFestimation in the absence of current measurement in the frames 3, 8 and 11. Bottom row shows the template matching results when the current measurement is ready. The object being tracked, the animal's faceis changing, since the face is turning. (This dataset was copied from www.vision.ucsd.edu/~bbabenko).

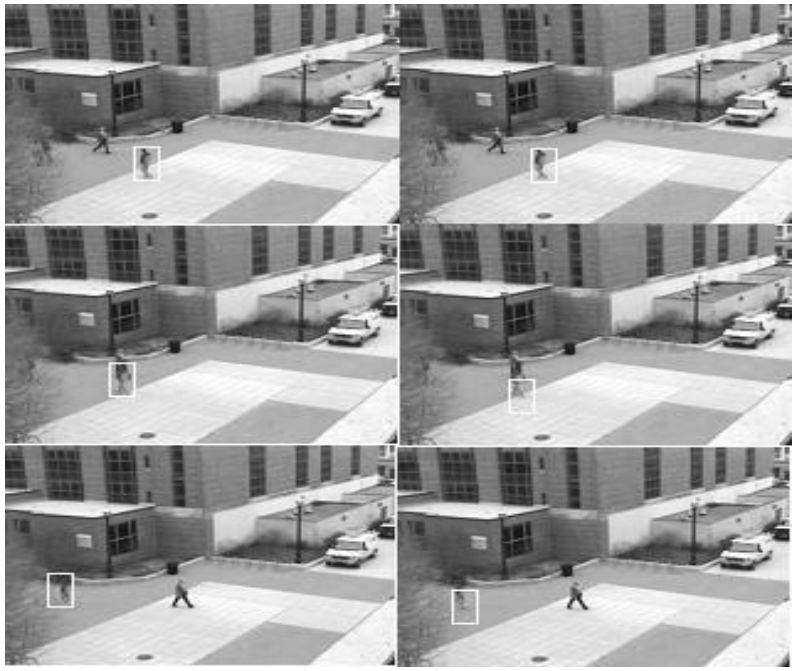


**Figure 10** Result similar to the previous Figures: the left column shows TCRF estimation without utilizing the current measurement in the frames 4, 6 and 12. The right column shows the template matching results. You can see in some situation like the second row, where the object is in occlusion condition, TCRF works as well as the previous. It is noteworthy that the frame number is assign in each shot.

## References

[1] Shafiee, M. J., Azimifar, Z., and Fieguth, P. (2010), "Model-based tracking: Temporal conditional random fields", ICIP.

[2] Taycher, L., Shakhnarovich, G., Demirdjian, D., and Darrell, T. (2006), "Conditional random people: Tracking humans with crfs and grid filters", CVPR, 222-229.

[3] Zhang, L., and Ji, Q. (2008), "Segmentation of video sequences using spatial-temporal conditional random fields", Proceedings of the 2008 IEEE Workshop on Motion and video Computing.

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

[5] Sigal, L., Zhu, Y., Comaniciu, D., and Black, D.(2007), "Tracking complex objects using graphicalobject models", Lecture Notes in Computer Science.

[6] Ablavsky,V., Thangali, A., and Sclaroff, S.(2008), "Layered graphical models for tracking partially-occluded objects" CVPR.

[7] Ren, X.(2008), "Finding people in archive films through tracking", CVPR.

[8] Boudoukh, G., Leichter, I., and Rivlin, E.(2009), "Visual tracking of object silhouettes", ICIP.

[9] Laerty, J., McCallum, A., and Pereira, F.(2001), "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", ICML.

[10] Klinge, R., and Tomanek, K.(2007), "Classical probabilistic models and conditional random fields algorithm engineering report", Tech. Rep., University of Manchester Stopford Building.

[11] Wallach, M.(2004), "Conditional random fields: An introduction", Tech. Rep., University of Pennsylvania CIS Technical Report MS-CIS-04-21.

[12]Taskar, B., Abbeel, P., and Koller, D.(2002), "Discriminative probabilistic models for relational data", In Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence.

[13]Sutton, C., McCallum, A., and Rohanimanesh, K.(2007), "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data", Journal of Machine Learning.

[14]Yin, J., Hu, D., and Yang, Q.(2009), "Spatio-temporal event detection using dynamic conditional random fields".

[15] McCallum, A., Rohanimanesh, K., and Sutton, C.(2003), "Dynamic conditional random fields for jointly labeling multiple sequences", nips workshop on syntax, semantics and statistics.

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

[17] Geman, S., and Geman, D.(1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", IEEE transaction on Pattern Recognition and Machine Intelligent.

[18] Wallach, H.(2002), "Efficient training of conditional random fields", Proc. 6th Annual CLUK Research Colloquium, Cite seer.

## *CCC Calls for New Manuscripts*

***Computer Communication & Collaboration*** is a peer-reviewed journal, published by Better Advances Press (BAP), sponsored by Academic Research Center of Canada (ARCC). This journal publishes research papers in the fields of general theories of computer science, computer communications, machine learning, data mining, intelligent collaboration and other relevant topics, both theoretical and empirical. Topics include, but are not limited to:

- Artificial intelligence
- Future Internet architecture, protocols and services
- Intelligent robotics
- Internet content search
- Machine learning
- Multimedia communication
- Network applications
- Network safety
- Next generation network
- On-line social networks

This journal is currently published in both printed and online versions. The published papers are free access and download from online databases. The publisher also provides the authors with many benefits, such as free PDFs, special subsidy from ARCC, academic sponsoring, and so on.

We are seeking submissions for forthcoming issues. All papers should be written in professional English. The length of 3000-8000 words is suggested. A paper template of accepted submission is available on our website www.bapress.ca. All manuscripts should be prepared in MS-word format, and sent to journal3c@gmail.com or ccc@bapress.ca , in **one way ONLY**.

**Please well record your date and way of paper submission.**

If your article is rejected after reviewed, the correspondence author will know this result within 7 **weeks** from the date of paper submission;

If your article is qualified after rigorous reviewing and finally published, it is expected to be published within 7 **months** from the date of paper submission.

(To continue on Cover 3rd)