sensors

MDPI

Article

Visual Object Tracking Based on Cross-Modality Gaussian-Bernoulli Deep Boltzmann Machines with RGB-D Sensors

Mingxin Jiang [1], Zhigeng Pan [2] and Zhenzhou Tang [3,*]

[1] Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huai'an 223003, China; jmx@hyit.edu.cn
[2] Digital Media &Interaction Research Center, Hangzhou Normal University, Hangzhou 310012, China; zgpan@hznu.edu.cn
[3] College of Physics and Electronic Information Engineering, Wenzhou University, Wenzhou 325035, China
* Correspondence: tzz@wzu.edu.cn

Academic Editor: Joonki Paik
Received: 1 December 2016; Accepted: 5 January 2017; Published: 10 January 2017

Abstract: Visual object tracking technology is one of the key issues in computer vision. In this paper, we propose a visual object tracking algorithm based on cross-modality featuredeep learning using Gaussian-Bernoulli deep Boltzmann machines (DBM) with RGB-D sensors. First, a cross-modality featurelearning network based on aGaussian-Bernoulli DBM is constructed, which can extract cross-modality features of the samples in RGB-D video data. Second, the cross-modality features of the samples are input into the logistic regression classifier, andthe observation likelihood model is established according to the confidence score of the classifier. Finally, the object tracking results over RGB-D data are obtained using aBayesian maximum a posteriori (MAP) probability estimation algorithm. The experimental results show that the proposed method has strong robustness to abnormal changes (e.g., occlusion, rotation, illumination change, etc.). The algorithm can steadily track multiple targets and has higher accuracy.

Keywords: Gaussian-Bernoulli deep Boltzmann machines; cross-modality features; Bayesian MAP; visual object tracking

1. Introduction

Visual object tracking is one of the key research topics in the field of computer vision. In recent years, it has had a wide range of applications, such as robot navigation, intelligent video surveillance, andvideo measurement [1–4]. Despite many research efforts, visual object tracking is still regarded as a challenging problem due to changes in object appearance, occlusions, complex motion, illumination variation and background clutter [5].

A typical visual object tracking algorithm often includes three major components: a state transition model, an observation likelihood model and a search strategy. A state transition model is used to model the temporal consistency of the states of a moving object, whereas an observation likelihood model describes the object and observations based on visual representations. Undoubtedly, feature representation is the most important factor in visual object tracking. Most of existing RGB-D trackers [6–8] tend to use hand-crafted features to represent target objects, such as Harr-like features [9], histogram of oriented gradients (HOG) [10], and local binary patterns (LBP) [11]. Hand-crafted features aim to describe some pre-defined image patterns, but they cannot capture thecomplex and specific characteristics of target objects. Hand-crafted features may lead to the loss of unrecoverable information which is suitable for tracking in different scenarios.

With the rapiddevelopment of computation power and the emergence of large-scale visual data, deep learning has received much attention and had a promising performance in computer vision tasks, e.g., object tracking [12], object detection [13], and image classification [14]. Wang et al. proposed a so-called deep learning tracker (DLT) for robust visual tracking [15]. DLT trackers learn generic features from auxiliary natural images offline. ADLT tracker cannot obtain deep features with temporal invariance, which is important for visual object tracking. In [16], the authors proposed a video tracking algorithm using learned hierarchical features in which the hierarchical features are learned via a two-layer convolutional neural network. Ding et al. [17] proposed a new tracking–learning–data architecture to transfer a generic object tracker to a blur invariant object tracker without deblurring image sequences. One of the research focuses of this paper is how to use deep learning effectively to extract the features of the target objects in RGB-D data.

To the best of our knowledge, the existing visual tracking methods using deep learning follow a similar procedure, which tracks objects in 2D sequences. Object tracking is performed over 2D video sequences in most early research works like TLD tracker [18], MIL tracker [19] and VTD tracker [20]. With the great popularity of affordable depth sensors, such as Kinect, Asus Xtion, and PrimeSense, an explosive growth of RGB-D data that can be used nowadays has been seen. Reliable depth images can provide valuable information to improve tracking performance. In [21], the author establishesa unified benchmark dataset of 100 RGB-D videos, which provide a foundation for further research in both RGB and RGB-D tracking. One of theresearch focuses of this paper is how to fuse RGB information and depth information effectively to improve the performance of visual object tracking in RGB-D data.

To overcome the problems in the existing methods, we propose a visual object tracking algorithm based on cross-modality feature learning using Gaussian-Bernoulli deep Boltzmann machines (DBM) over RGB-D data. A cross-modality deep learning framework is usedto learn a robust tracker forRGB-D data. The cross-modality features of the samples are input into the logistic regression classifier, andthe observation likelihood model is established according to the confidence score of the classifier. We obtain the object tracking results over RGB-D data using aBayesian maximum a posteriori probability estimation algorithm. Experimental results show that such a cross-modality learning can improve the tracking performance.

The main contributions of this paper can be summarized as follows:

- We present a cross-modality Gaussian-Bernoulli deep Boltzmann machine (DBM) to learn the cross-modality features of target objects in RGB-D data. The proposed cross-modality Gaussian-Bernoulli DBM is constructed with two single-modality Gaussian-Bernoulli DBMs by adding an additional layer of binary hidden units on top of them, which can fuse RGB information and depth information effectively.
- A unified RGB-D tracking framework based on Bayesian MAP is proposed, in which the robust appearance description with cross-modality features deep learning, temporal continuity is fully considered in the state transition model.
- Extensive experiments are conducted to compare our tracker with several state-of-the-art methods on the recent benchmark dataset [21]. From experimental results, we can see that the proposed tracker performs favorably against the compared state-of-the-art trackers.

The remainder of the paper is organized as follows. First, feature learning over RGB-D data with cross-modality deep Boltzmann machines is described in the next section. Then we introduce our tracking framework in Section 3. The implementation of our proposed method is presented in Section 4. Experimental results and analysis are demonstrated in Section 5, and finally we draw conclusions in Section 6.

2. Related Work

2.1. Boltzmann Machine

TheBoltzmann machine (BM) was proposed by Hinton and Sejnowski [22]. A Boltzmann machine is a feedback neural network consisting of fully connected coupled random neurons. The connections between neurons are symmetric, and there is no self-feedback. The outputs of neurons only have two states (active and inactive) which are expressed by 0 and 1, respectively. A set of visible units $\mathbf{v} \in \{0,1\}^D$ and a set of hidden units $\mathbf{h} \in \{0,1\}^F$ are included in BM (as shown in Figure 1). The visible units and hidden units are composed ofthe visible nodes and hidden nodes, and D and F represent the number of visible nodes and hidden layer nodes, respectively.

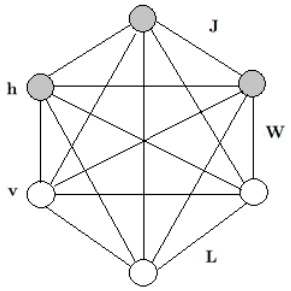

Figure 1. A general Boltzmann machine.

We formulate the energy function over the state $\{\mathbf{v}, \mathbf{h}\}$ as:

$$E(\mathbf{v}, \mathbf{h}; \Psi) = -\mathbf{v}'\mathbf{W}\mathbf{h} - \frac{1}{2}\mathbf{h}'\mathbf{R}\mathbf{h} - \frac{1}{2}\mathbf{v}'\mathbf{L}\mathbf{v} - \mathbf{v}'\mathbf{B} - \mathbf{h}'\mathbf{A} \tag{1}$$

where $\Psi = \{\mathbf{W}, \mathbf{L}, \mathbf{R}, \mathbf{B}, \mathbf{A}\}$ are the model parameters: $\mathbf{W}, \mathbf{L}, \mathbf{R}$ represent the symmetric interaction terms of visible nodes to hidden nodes, visible nodes to visible nodes, and hidden nodes to hidden nodes. The diagonal elements of $\mathbf{L}$ and $\mathbf{R}$ are set to 0. $\mathbf{B}$ and $\mathbf{A}$ are the threshold values of the visible layer and the hidden layer.

The model defines a probability distribution over a visible vector $\mathbf{v}$ as:

$$P(\mathbf{v}; \Psi) = \frac{P^*(\mathbf{v}; \Psi)}{Z(\Psi)} = \frac{1}{Z(\Psi)} \sum_h exp(-E(\mathbf{v}, \mathbf{h}; \Psi)) \tag{2}$$

where $Z(\Psi) = \sum_{\mathbf{v}} \sum_{\mathbf{h}} exp(-E(\mathbf{v}, \mathbf{h}; \Psi))$ is called the partition function, and P^* is an unnormalized probability.

The following formulations give the conditional distributions over hidden and visible units:

$$P(h_j = 1 | \mathbf{v}, \mathbf{h}_{-j}) = \sigma(\sum_{i=1}^{D} W_{ij} v_i + \sum_{m=1 \backslash j}^{P} J_{jm} h_j) \tag{3}$$

$$P(v_i = 1 | \mathbf{h}, \mathbf{v}_{-i}) = \sigma(\sum_{j=1}^{P} W_{ij} h_j + \sum_{k=1 \backslash i}^{D} J_{ik} v_i) \tag{4}$$

where $\sigma(x) = 1/(1 + exp(-x))$ is the logistic function.

2.2. Restricted Boltzmann Machine

Setting both $\mathbf{L} = 0$ and $\mathbf{R} = 0$ in Equation (1), we will recover the model of a restricted Boltzmann machine (RBM), as shown in Figure 2.

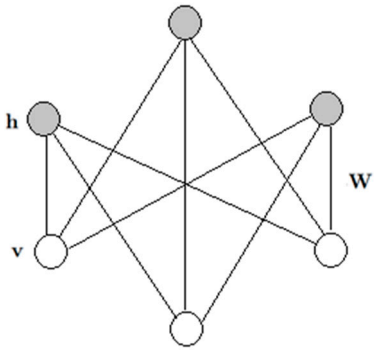

Figure 2. A restricted Boltzmann machine.

A restricted Boltzmann machine(RBM) is a generative stochastic artificial neural networkthat can learn a probability distribution over its set of inputs. It is an undirected graphical model with each visible unit only connected to each hidden unit. The energy function over the visible and hidden units.

$$E(\mathbf{v}, \mathbf{h}; \Psi) = -\mathbf{v}'\mathbf{W}\mathbf{h} - \mathbf{v}'\mathbf{B} - \mathbf{h}'\mathbf{A} \tag{5}$$

where $E : \{0,1\}^{D+F} \rightarrow \mathbb{R}$, $\Psi = \{\mathbf{W}, \mathbf{A}, \mathbf{B}\}$ are the model parameters. Equation (6) defines the joint probability distribution over the visible units $\mathbf{v} \in \{0,1\}^D$ and hidden units $\mathbf{h} \in \{0,1\}^F$.

$$P(\mathbf{v}, \mathbf{h}; \Psi) = \frac{1}{Z(\Psi)} exp(-E(\mathbf{v}, \mathbf{h}; \Psi)) \tag{6}$$

where the normalizing factor $Z(\Psi)$ denotes the partition function.

2.3. Gaussian-Bernoulli Restricted Boltzmann Machines

When inputs are real-valued images, we formulate the energy function of the Gaussian-Bernoulli RBM over the state $\{\mathbf{v}, \mathbf{h}\}$ as follows [23]:

$$E(\mathbf{v}, \mathbf{h}; \Psi) = -\sum_{i=1}^{D} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i=1}^{D}\sum_{j=1}^{F} W_{ij} h_j \frac{v_i}{\sigma_i} - \sum_{j=1}^{F} a_j h_j \tag{7}$$

where $\Psi = \{\mathbf{a}, \mathbf{b}, \mathbf{W}, \sigma\}$ are the model parameters, b_i and a_j are biases corresponding to visible and hidden variables, respectively, W_{ij} is the matrix of weights connecting visible and hidden nodes, and σ_i is the standard deviation associated with a Gaussian visible variable v_i.

2.4. Gaussian-Bernoulli Deep Boltzmann Machine

A deep Boltzmann machine (DBM) [24] contains a set of visible units $\mathbf{v} \in \{0,1\}^D$, and a sequence of layers of hidden units $\mathbf{h}^1 \in \{0,1\}^{L_1}$, $\mathbf{h}^2 \in \{0,1\}^{L_2}$, ... , $\mathbf{h}^N \in \{0,1\}^{L_N}$. Connections only exist between hidden units in adjacent layers. We illustrate a two-layer Gaussian-Bernoulli deep Boltzmann machine, consisting of learning a stack of modified Gaussian-Bernoulli RBMs (see Figure 3).

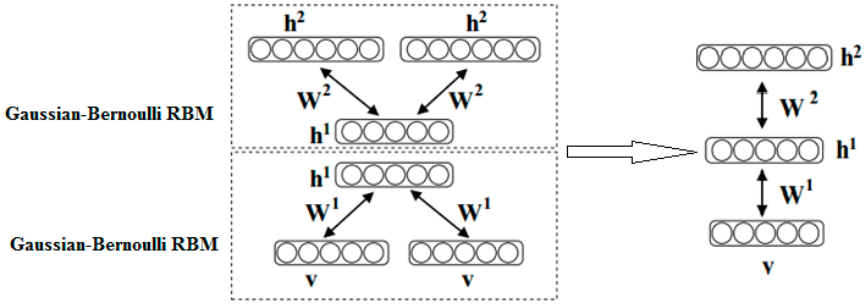

Figure 3. A Gaussian-Bernoulli Deep Boltzmann Machine.

The energy function of the joint configuration $\{\mathbf{v}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}\}$ is formulated as:

$$E(\mathbf{v}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}; \Psi) = -\mathbf{v}'\mathbf{W}^{(1)}\mathbf{h}^{(1)} - \mathbf{h}^{(1)'}\mathbf{W}^{(2)}\mathbf{h}^{(2)} \tag{8}$$

where $\Psi = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}\}$ are the model parameters, and $\mathbf{h} = \{\mathbf{h}^{(1)}, \mathbf{h}^{(2)}\}$ denote the set of hidden units. The probability distribution over a visible vector $\mathbf{v}$ can be modelled as:

$$P(\mathbf{v}; \Psi) = \frac{1}{Z(\Psi)} \sum_{\mathbf{h}^{(1)}, \mathbf{h}^{(2)}} exp(-E(\mathbf{v}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}; \Psi)) \tag{9}$$

3. Proposed Tracking framework

3.1. Feature Learning UsingCross-Modality Deep Boltzmann Machines over RGB-D Data

ABoltzmann machine (BM) is an effective tool in representing probability distribution over its inputs. Deep Boltzmann Machines (DBMs) have been successfully used in many application domains, e.g., topic modelling, classification, dimensionality reduction, feature learning, etc. According to the task, DBMs can be trained in either unsupervised or supervised ways. In this paper, we propose the cross-modality DBMs for feature learning in visual tracking over RGB-D data. In this section, we first describe how to establish cross-modality DBMs, review BMs, RBMs and Gaussian-Bernoulli restricted Boltzmann machines, then go over them in detail.

Multimodal deep learning was proposed forvideo and audio [25,26]. In RGB-D data, we can also learn deep features over multiple modalities (RGB modality and depth modality). The proposed cross-modality DBM is constructed with two single-modality Gaussian-Bernoulli DBMs by adding an additional layer of binary hidden units on top of them (see Figure 4). Firstly, we model a RGB-specific Gaussian-Bernoulli DBM with two hidden layers as Figure 4a, where $\mathbf{v}^{RGB} \in \mathbb{R}^D$ denotes a real-valued image input. Let $\mathbf{h}^{(1RGB)} \in \{0,1\}^{F_1^{RGB}}$ and $\mathbf{h}^{(2RGB)} \in \{0,1\}^{F_2^{RGB}}$ be the two layers of hidden units in the RGB-specific DBM. Then, the energy function of Gaussian-Bernoulli DBM over $\{\mathbf{v}^{RGB}, \mathbf{h}^{RGB}\}$ is defined as:

$$\begin{aligned} E(\mathbf{v}^{RGB}, \mathbf{h}^{(1RGB)}, \mathbf{h}^{(2RGB)}; \Psi^{RGB}) = & \sum_{i=1}^{D} \frac{(v_i^{(RGB)} - b_i^{(RGB)})^2}{2\sigma_i^{(RGB)^2}} - \sum_{i=1}^{D} \sum_{j=1}^{F_1^{RGB}} \frac{v_i^{(RGB)}}{\sigma_i^{(RGB)}} W_{ij}^{(1RGB)} h_j^{(1RGB)} \\ & - \sum_{j=1}^{F_1^{RGB}} \sum_{l=1}^{F_2^{RGB}} W_{jl}^{(2RGB)} h_j^{(1RGB)} h_l^{(2RGB)} - \sum_{j=1}^{F_1^{RGB}} a_j^{(1RGB)} h_j^{(1RGB)} - \sum_{l=1}^{F_2^{RGB}} a_l^{(2RGB)} h_l^{(2RGB)} \end{aligned} \tag{10}$$

where $\sigma_i^{(RGB)}$ is thedeviation of the corresponding Gaussian model, and Ψ^{RGB} is the parameter vector of RGB-specific Gaussian-Bernoulli DBM. Therefore, the joint distribution of the energy-based probabilistic model is defined through an energy function as:

$$P(\mathbf{v}^{RGB}, \mathbf{h}^{RGB}; \Psi^{RGB}) = \frac{1}{Z(\Psi^{RGB})} \sum_{\mathbf{h}^{RGB}} exp(-E(\mathbf{v}^{RGB}, \mathbf{h}^{RGB}; \Psi^{RGB})) \tag{11}$$

where $Z(\Psi^{RGB})$ is the partition function.

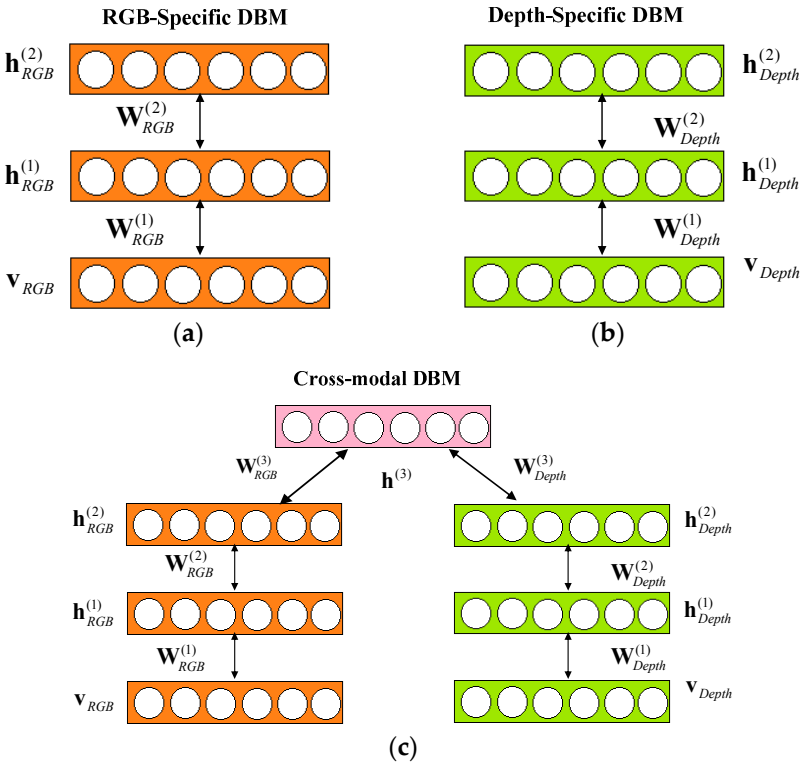

Figure 4. The illustration of the proposed cross-modal Gaussian-Bernoulli DBM. (**a**) RGB-specific two-layer Gaussian-Bernoulli DBM; (**b**) Depth-specific two-layer Gaussian-Bernoulli DBM; (**c**) a Cross-modal Gaussian-Bernoulli DBM.

Similarly, the corresponding probability assigned to $\mathbf{v}^{Depth}$ by Depth-specific DBM has the same form with Equation (11). Let $\mathbf{v}^{Depth} \in \mathbb{R}^K$ denotes a real-valued depth image input. Let $\mathbf{h}^{(1Depth)} \in \{0,1\}^{F_1^{Depth}}$ and $\mathbf{h}^{(2Depth)} \in \{0,1\}^{F_2^{Depth}}$ be the two layers of hidden units in the Depth-specific DBM, as show in Figure 4b. The energy of the Gaussian-Bernoulli DBM and the joint distribution of the energy-based probabilistic model over $\left\{\mathbf{v}^{Depth}, \mathbf{h}^{Depth}\right\}$ are defined as:

$$\begin{aligned} E(\mathbf{v}^{Depth}, \mathbf{h}^{(1Depth)}, \mathbf{h}^{(2Depth)}; \Psi^{Depth}) = & \sum_{i=1}^{D} \frac{(v_i^{(Depth)} - b_i^{(Depth)})^2}{2\sigma_i^{(Depth)^2}} - \sum_{i=1}^{D} \sum_{j=1}^{F_1^{Depth}} \frac{v_i^{(Depth)}}{\sigma_i^{(Depth)}} W_{ij}^{(1Depth)} h_j^{(1Depth)} \\ & - \sum_{j=1}^{F_1^{Depth}} \sum_{l=1}^{F_2^{Depth}} W_{jl}^{(2Depth)} h_j^{(1Depth)} h_l^{(2Depth)} - \sum_{j=1}^{F_1^{Depth}} a_j^{(1Depth)} h_j^{(1Depth)} - \sum_{l=1}^{F_2^{Depth}} a_l^{(2Depth)} h_l^{(2Depth)} \end{aligned} \tag{12}$$

$$P(\mathbf{v}^{Depth}, \mathbf{h}^{(1Depth)}, \mathbf{h}^{(2Depth)}; \Psi^{Depth}) = \frac{1}{Z(\Psi^{Depth})} \sum_{\mathbf{h}^{(1Depth)(2Depth)}} exp(-E(\mathbf{v}^{Depth}, \mathbf{h}^{(1Depth)}, \mathbf{h}^{(2Depth)}; \Psi^{Depth})) \tag{13}$$

where $\sigma_i^{(Depth)}$ is deviation of the corresponding Gaussian model, and Ψ^{Depth} is the parameter vector of Depth-specific Gaussian-Bernoulli DBM.

Let $\mathbf{v}^{RGB} \in \mathbb{R}^D$ and $\mathbf{v}^{Depth} \in \mathbb{R}^K$ denote a real-valued RGB input and a real-valued depth input respectively. Consider modeling an image-depth DBM with three hidden layers, let $\{\mathbf{v}^{RGB}, \mathbf{v}^{Depth}\}$ be real-valued Gaussian variables, and $\{\mathbf{h}^{(1RGB)}, \mathbf{h}^{(2RGB)}, \mathbf{h}^{(1Depth)}, \mathbf{h}^{(2Depth)}, \mathbf{h}^{(3)}\}$ be binary stochastic hidden units. Let $\mathbf{h}^{(1RGB)} \in \{0,1\}^{F_1^{RGB}}$ and $\mathbf{h}^{(2RGB)} \in \{0,1\}^{F_2^{RGB}}$ be the two layers of hidden units in the RGB-specific two layer DBM. Similarly, let $\mathbf{h}^{(1Depth)} \in \{0,1\}^{F_1^{Depth}}$ and $\mathbf{h}^{(2Depth)} \in \{0,1\}^{F_2^{Depth}}$ be the two layers of hidden units in the depth-specific two layer DBM. The energy of the proposed cross-modality Gaussian-Bernoulli DBM over $\{\mathbf{v}, \mathbf{h}\}$ can be defined as:

$$\begin{aligned}
&E(\mathbf{v},\mathbf{h};\Psi^{cross-modality}) = \sum_{i=1}^{D}\frac{(v_i^{(RGB)}-b_i^{(RGB)})^2}{2\sigma_i^{(RGB)^2}} - \sum_{i=1}^{D}\sum_{j=1}^{F_1^{RGB}}\frac{v_i^{(RGB)}}{\sigma_i^{(RGB)}}W_{ij}^{(1RGB)}h_j^{(1RGB)} \\
&- \sum_{j=1}^{F_1^{RGB}}\sum_{l=1}^{F_2^{RGB}} h_j^{(1RGB)}W_{jl}^{(2RGB)}h_l^{(2RGB)} - \sum_{l=1}^{F_2^{RGB}}\sum_{p=1}^{F_3^{RGB}} h_j^{(2RGB)}W_{lp}^{(3RGB)}h_p^{(3RGB)} - \sum_{j=1}^{F_1^{RGB}} a_j^{(1RGB)}h_j^{(1RGB)} \\
&- \sum_{l=1}^{F_2^{RGB}} a_l^{(2RGB)}h_l^{(2RGB)} + \sum_{i=1}^{K}\frac{(v_i^{(Depth)}-b_i^{(Depth)})^2}{2\sigma_i^{(\mathrm{Depth})^2}} - \sum_{i=1}^{K}\sum_{j=1}^{F_1^{Depth}}\frac{v_i^{(Depth)}}{s_i^{(Depth)}}W_{ij}^{(1Depth)}h_j^{(1Depth)} \\
&- \sum_{j=1}^{F_1^{Depth}}\sum_{l=1}^{F_2^{Depth}} h_j^{(1Depth)}W_{jl}^{(2Depth)}h_l^{(2Depth)} - \sum_{l=1}^{F_2^{Depth}}\sum_{p=1}^{F_3^{Depth}} h_j^{(2Depth)}W_{lp}^{(3Depth)}h_p^{(3Depth)} - \sum_{j=1}^{F_1^{Depth}} a_j^{(1Depth)}h_j^{(1Depth)} \\
&- \sum_{l=1}^{F_2^{Depth}} a_l^{(2Depth)}h_l^{(2Depth)} - \sum_{p=1}^{F_3} a_p^{(3)}h_p^{(3)}
\end{aligned} \tag{14}$$

Therefore, the joint probability distribution over the cross-modal input $\{\mathbf{v}^{RGB}, \mathbf{v}^{Depth}\}$ can be written as:

$$\begin{aligned}
&P(v^{RGB}, v^{Depth};\Psi^{cross-modality}) = \sum_{h^{(2RGB)},h^{(2Depth)},h^{(3)}} P(h^{(2RGB)},h^{(2Depth)},h^{(3)})\Big(\sum_{h^{(1RGB)}} P(v^{RGB},h^{(1RGB)},h^{(2RGB)})\Big) \\
&\Big(\sum_{h^{(1Depth)}} P(v^{Depth},h^{(1Depth)},h^{(2Depth)})\Big) \\
&= \frac{1}{Z(\Psi^{cross-modality})}\sum_h exp\Big(-\sum_i\frac{(v_i^{RGB})^2}{2s_i^2} + \sum_{ij}\frac{v_i^{(RGB)}}{s_i}W_{ij}^{(1RGB)}h_j^{(1RGB)} + \sum_{jl}W_{jl}^{(2RGB)}h_j^{(1RGB)}h_l^{(2RGB)} \\
&-\sum_i\frac{(v_i^{Depth})^2}{2s_i^2} + \sum_{ij}\frac{v_i^{(Depth)}}{s_i}W_{ij}^{(1Depth)}h_j^{(1Depth)} + \sum_{jl}W_{jl}^{(2Depth)}h_j^{(1Depth)}h_l^{(2Depth)} \\
&+\sum_{lp}W^{(3RGB)}h_l^{(2RGB)}h_p^{(3)} + \sum_{lp}W^{(3Depth)}h_l^{(2Depth))}h_p^{(3)}\Big)
\end{aligned} \tag{15}$$

where $\Psi^{cross-modality}$ is the parameter vector of cross-modality Gaussian-Bernoulli DBM. The task of learning the cross-modality Gaussian-Bernoulli DBM is the maximum likelihood learning for Equation (6) with respect to the model parameters.

3.2. Bayesian Framework

In this paper, the object tracking is formulated as a hidden state variable Bayesian maximum a posteriori (MAP) estimation problem in the Hidden Markov model. Given a set of observed variables $\mathbf{Z}_t = \{Z_1, Z_2, \ldots, Z_t\}$, we can estimate the hidden state variable $\mathbf{X}_t = \{\mathbf{X}_t^1, \mathbf{X}_t^2, \ldots\ldots \mathbf{X}_t^N\}$ by using Bayesian MAP theory [27].

The posteriori probability distribution according to the Bayesian theory can be modelled as the following derivation:

$$p(\mathbf{X}_t|\mathbf{Z}_t) \propto p(\mathbf{Z}_t|\mathbf{X}_t)\int p(\mathbf{X}_t|\mathbf{X}_{t-1})p(\mathbf{X}_{t-1}|\mathbf{Z}_{t-1})d\mathbf{X}_{t-1} \tag{16}$$

where $p(\mathbf{Z}_t|\mathbf{X}_t)$ stands for an observation likelihood model and $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ is called a state transition model for two consecutive frames. We can obtain the optimal state $\hat{\mathbf{X}}_t$ among all the candidates through maximum posterior probability estimation:

$$\hat{\mathbf{X}}_t = \arg\max_{\mathbf{X}_t} p(\mathbf{X}_t|\mathbf{Z}_t) \tag{17}$$

3.2.1. State Transition Model

The state variable is defined as $\mathbf{X}_t = \{x_t, y_t, \theta_t, s_t, \alpha_t, \phi_t\}$, which includes the six parameters of the motion affine transformation, where x_t and y_t denote the x-direction and y-direction translation of the object in the frame t respectively, θ_t represents the rotation angle, s_t stands for the scale change, α_t denotes the aspect ratio, and ϕ_t represents skew direction at time t.

We assume that the candidate states are generated according to Gaussian distribution:

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}) \ = \ N(\mathbf{X}_t; \mathbf{X}_{t-1}, \Sigma) \tag{18}$$

where Σ is a diagonal covariance matrix whose diagonal elements are $\sigma_x^2, \sigma_y^2, \sigma_\theta^2, \sigma_S^2, \sigma_\alpha^2, \sigma_\phi^2$.

3.2.2. Observation Likelihood Model

In this paper, the observation model that we use is discriminative. A binary linear classifier is adopted to classify tracking observations into object class and background class during tracking. Observations are represented using features learned from the DBM introduced previously. We can obtain a training dataset with approximate labels after extracting features of positive and negative samples. Deep representations are likely to be linearly separable, and linear classifiers are less prone to overfitting. We adopt the logistic regression classifier owing to its capability of providing predictions in probability estimation.

Let $\mathbf{h}_i^3 \in \mathbb{R}^{r\times 1}$ denote the deep feature for the i-th training sample, and $y_i \in \{-1, +1\}$ represent the label for the i-th training sample. $Z^+ = [\mathbf{h}_{1^+}^3, \mathbf{h}_{2^+}^3, \ldots, \mathbf{h}_{D^+}^3] \in \mathbb{R}^{r\times D^+}$ stands for the positive training set with their respective labels as $Y^+ = [y_{1^+}, y_{2^+}, \ldots, y_{D^+}] \in \{-1, +1\}^{D^+\times 1}$. Similarly, $Z^- = [\mathbf{h}_{1^-}^3, \mathbf{h}_{2^-}^3, \ldots, \mathbf{h}_{D^-}^3] \in \mathbb{R}^{r\times D^-}$ represents the negative training set with their respective labels as $Y^- = [y_{1^-}, y_{2^-}, ..., y_{D^-}] \in \{-1, +1\}^{D^-\times 1}$. Training the logistic regression classifier by optimizing:

$$\min_{\pm w} C^+ \sum_{i^+=1}^{D^+} \log(1 + e^{y_{i^+} \pm w^T \pm h_{i^+}^{(3)}}) + C^- \sum_{i^-=1}^{D^-} \log(1 + e^{y_{i^-} \pm w^T \pm h_{i^-}^{(3)}}) \tag{19}$$

where $C^+ \in \mathbb{R}$ is the parameter to weight the logistic cost of the positive-class and $C^- \in \mathbb{R}$ is the parameter to weight the logistic cost of the negative-class logistic. Weight regularization w is added to the cost function in Equation (19) to reduce overfitting. In the prediction stage, the confidence score of the trained logistic regression classifier can be computed as follows:

$$p(\mathbf{Z}_t|\mathbf{X}_t) = \frac{1}{1 + e^{-(\pm w^T \pm \mathbf{z}_t)}} \tag{20}$$

4. The Implementation of Our Proposed Method

Our method has two major components, which are shown in Figures 5 and 6. In the first place, as demonstrated in Figure 5, unlabeled patches in RGB and depth modality are used to train the cross-modality Gaussian-Bernoulli DBM offline.

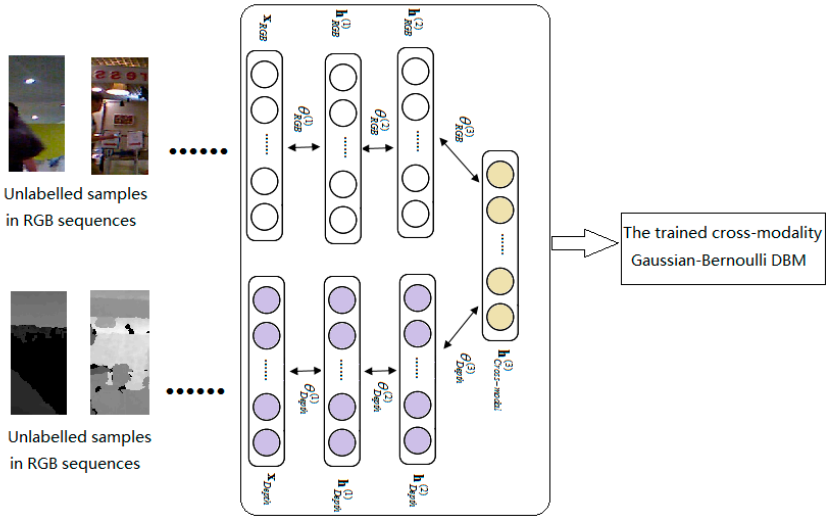

Figure 5. Offline learning of the proposed cross-modality Gaussian-Bernoulli DBM.

Then, the trained cross-modality Gaussian-Bernoulli DBM is transferred to an observational model for visual tracking online based on Bayesian MAP, as shown in Figure 6.

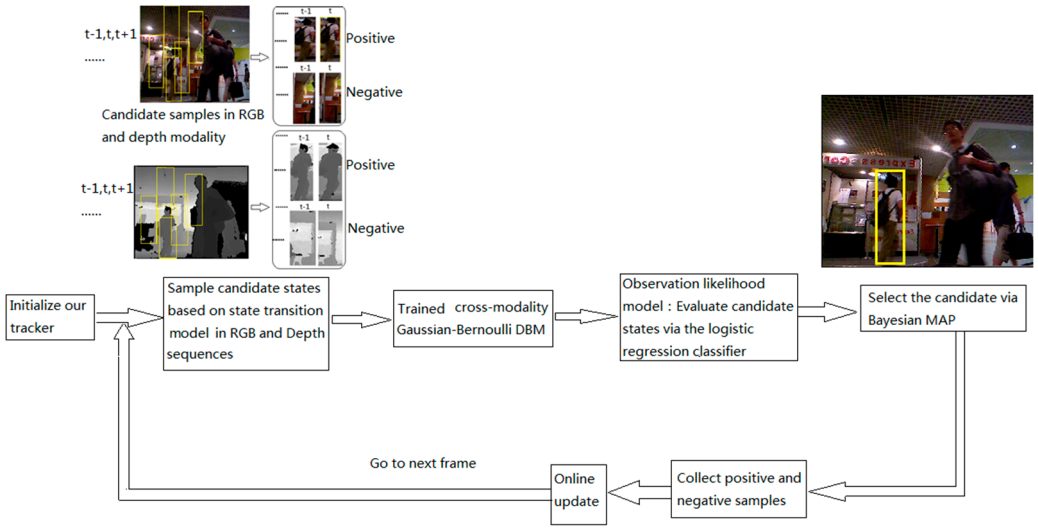

Figure 6. The process of object tracking online based on Bayesian MAP.

5. Experimental Results and Analysis

The experiments of our proposed tracking algorithm is implemented on MATLAB R2014a, Intel(R) Core(TM) i7-4712MQ, CPU@3.40 GHz and TITAN GPU, 8.00 GB RAM, Windows 8.1 operating system, in Beijing, China.

5.1. Qualitative Evaluation

In order to show the robustness of the visual object tracking algorithm discussed in this paper, we compare our tracker with several state-of-the-art methods on arecent benchmark dataset [21] in different environments with heavy or long-time partial occlusion, rotation, scale change, and fast motion. Given the limited space, in this section we only list four of them to show the experimental results and the forms of data statistics.

We compare our method with several state-of-the-art trackers, including TLD Tracker[18], MIL Tracker [19],VTD Tracker [20], and RGB-D Tracker [28], CT Tracker [29], Struck Tracker [30], Deep Tracker [15], and Multi-cues Tracker [31],andwe ranthe experiments based on the code provided by the authors.

Figure 7 demonstrates that our method performs well in terms of rotation, scale and position when the object undergoes severe occlusion. The MIL tracker and VTD tracker are sensitive to occlusion.

Figure 7. The tracking results on the test video 1 obtained by different methods.

Figure 8 shows the tracking results in the sequence with long-time partial occlusion, pose change and background clutter. We can see that the RGBD, MIL and VTD methods do not perform well and they are less effective in this case.

Figure 8. *Cont.*

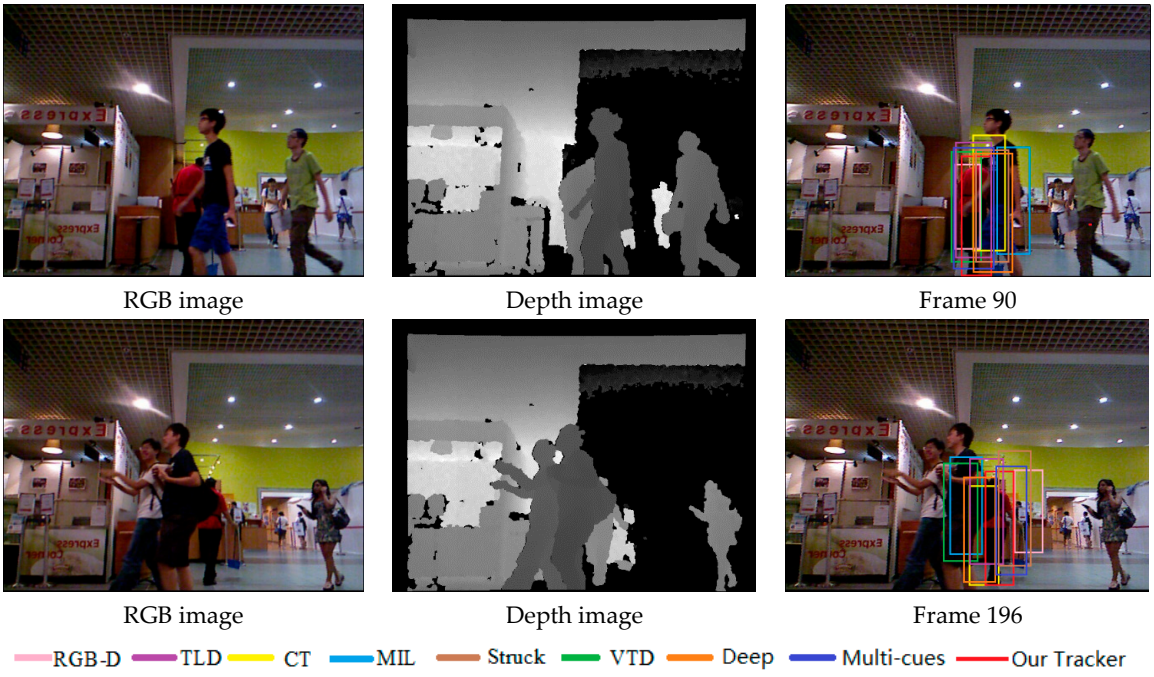

Figure 8. The tracking results on the test video 2 obtained by different methods.

Figure 9 illustrates the tracking results on the test video with severe occlusion, appearance change and fast motion. From the results, we can notice that the TLD, MIL and VTD methods are sensitive to target appearance change or occlusion.

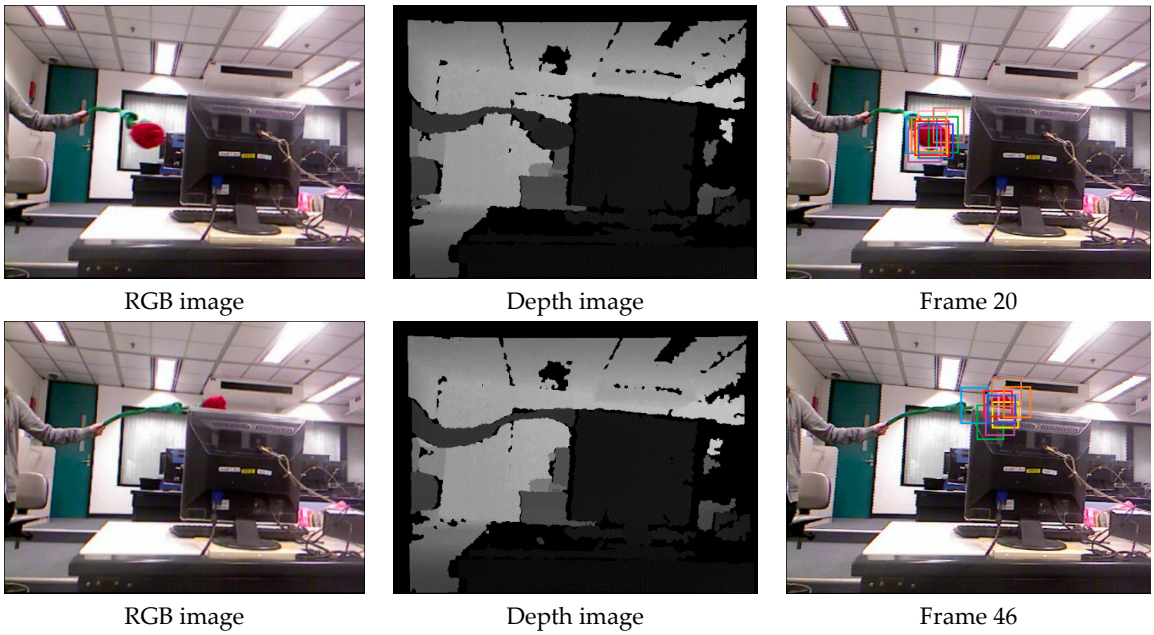

Figure 9. *Cont.*

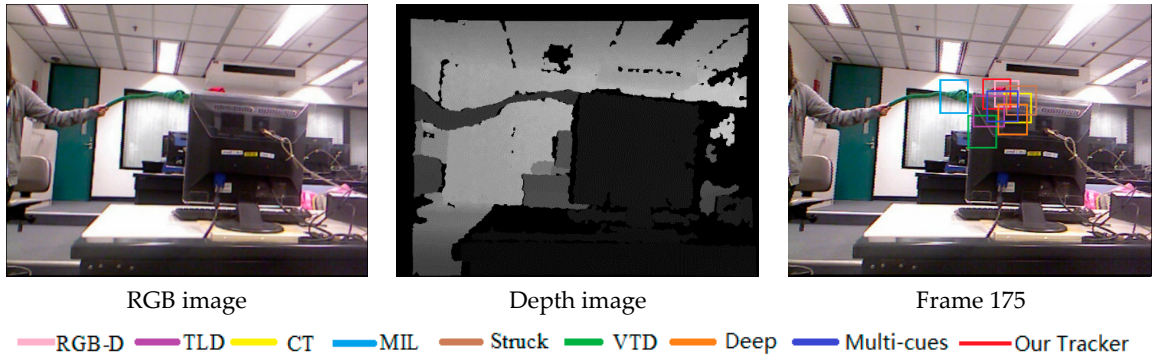

Figure 9. The tracking results on the test video 3 obtained by different methods.

Figure 10 shows the tracking results in the sequence with all occlusion, pose change and background clutter. We can see that the Stuck, MIL and VTD methods do not perform well and they are less effective in this case.

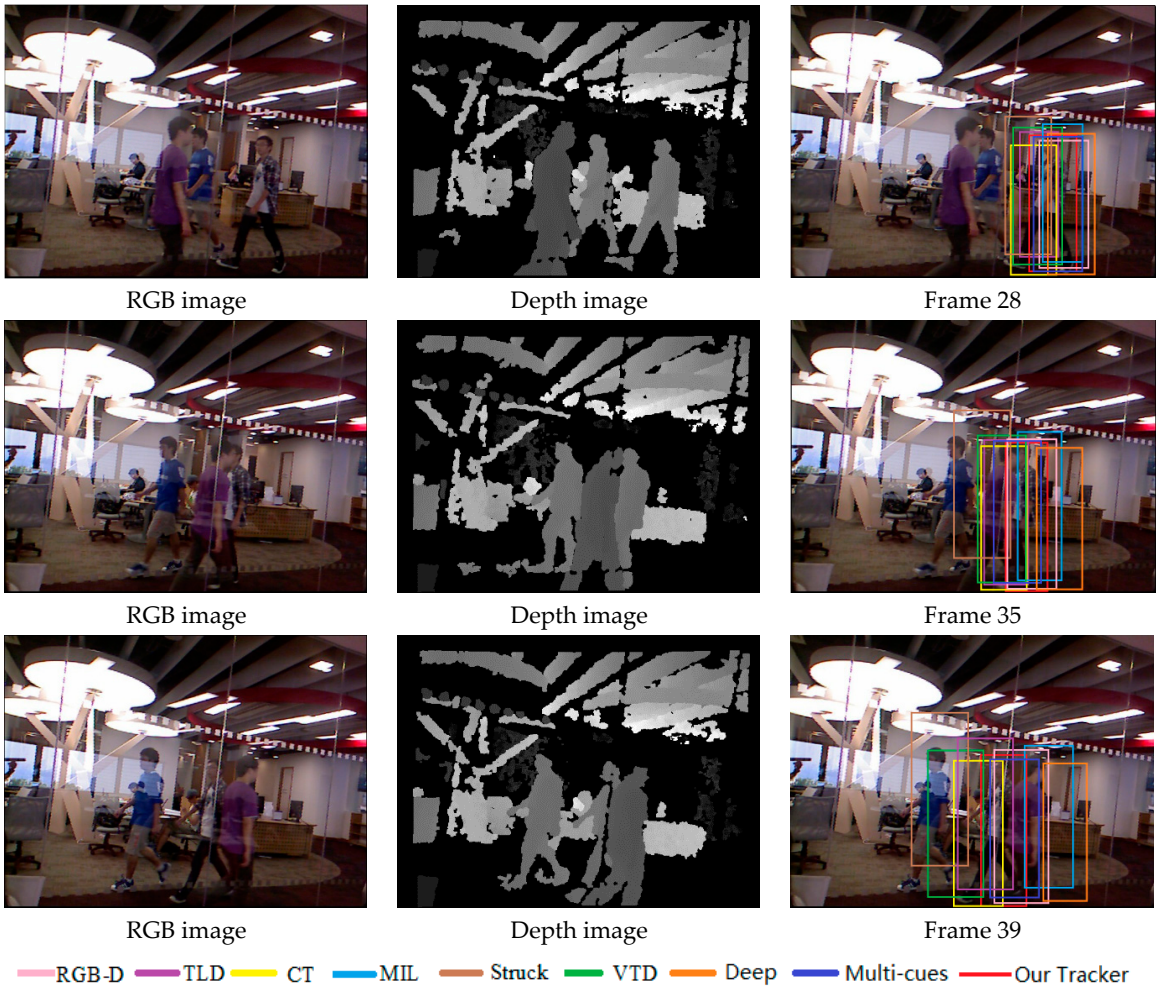

Figure 10. The tracking results on the test video 4 obtained by different methods.

Figure 11 illustrates the "bad" tracking results of our method,meaning frames where tracking failures are observed. When the objects are all occluded, the tracking results of our method experience a drift phenomenon.

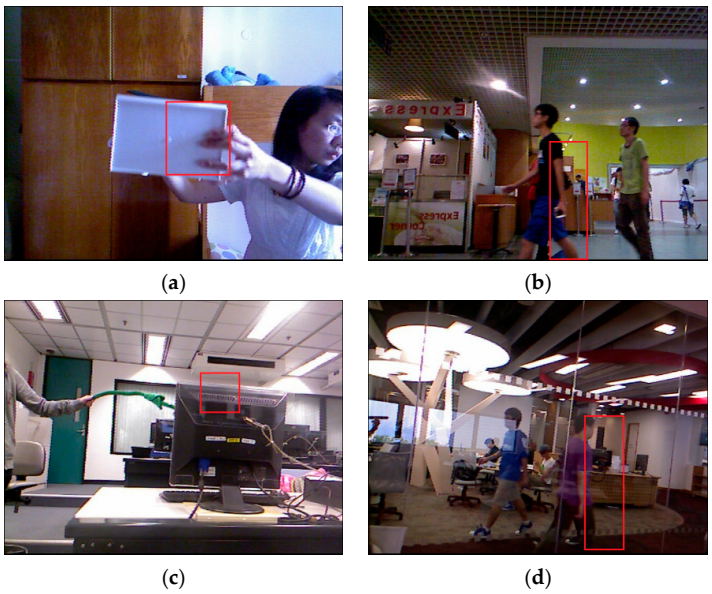

Figure 11. The"bad" tracking results of our method. (**a**) Frame 47 in test video 1; (**b**) Frame 93 in test video 1; (**c**) Frame 66 in test video 3; (**d**) Frame 37 in test video 3.

As shown in experimental results, the proposed tracking method performs favorably against the state-of-the-art tracking methods in handling challenging video sequences, but there are some limitations for our method. The robustness of the proposed tracking method is not strong enough to solve allocclusion and abrupt movement.

5.2. Quantitative Evaluation

We use two measurements to quantitatively evaluate tracking performances. The first one is called average center location error [32] which measures distances of centers between tracking results and ground truths in pixels. The second one is called success rate (SR) which is calculated according to $\frac{area(R_T \cap R_G)}{area(R_T \cup R_G)}$ and indicates theextent of region overlapping between tracking results R_T and R_G.

Figures 12–15 report the average center location errors of different tracking methods over three test videos. The comparison results show that the proposed method has a smaller average center location error than the state-of-the-art methods indifferent situations.

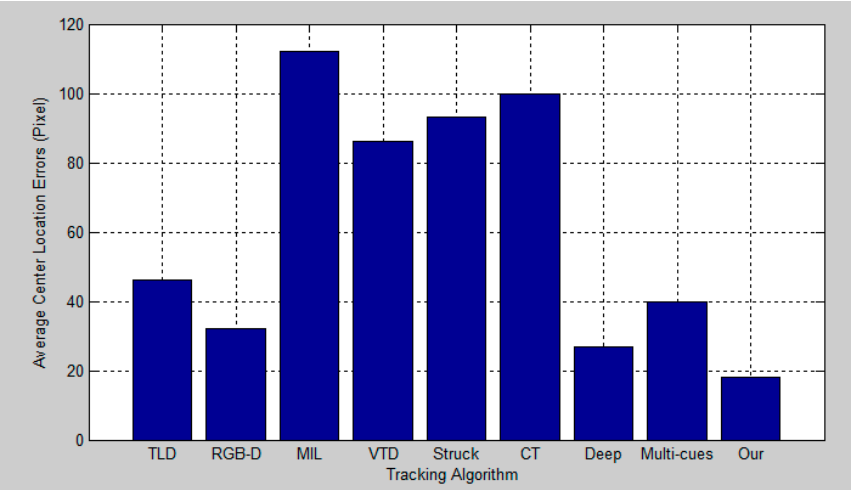

Figure 12. Quantitative evaluation in terms of average center location error (in pixel) for the first experiment.

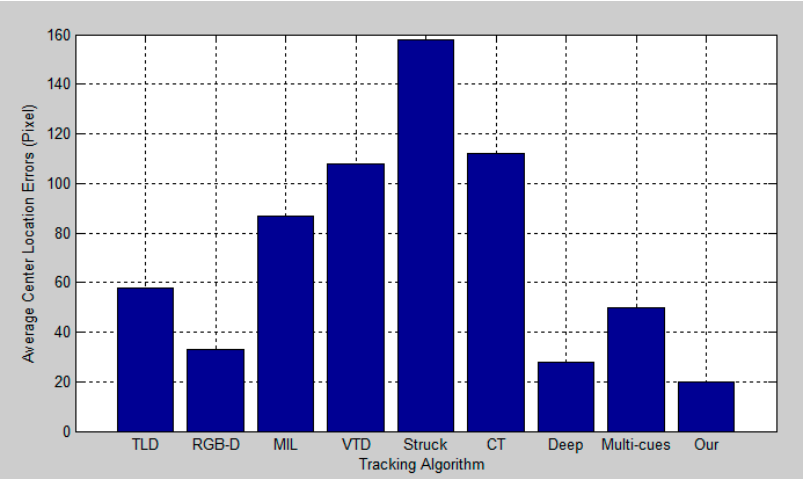

Figure 13. Quantitative evaluation in terms of average center location error (in pixel) for the second experiment.

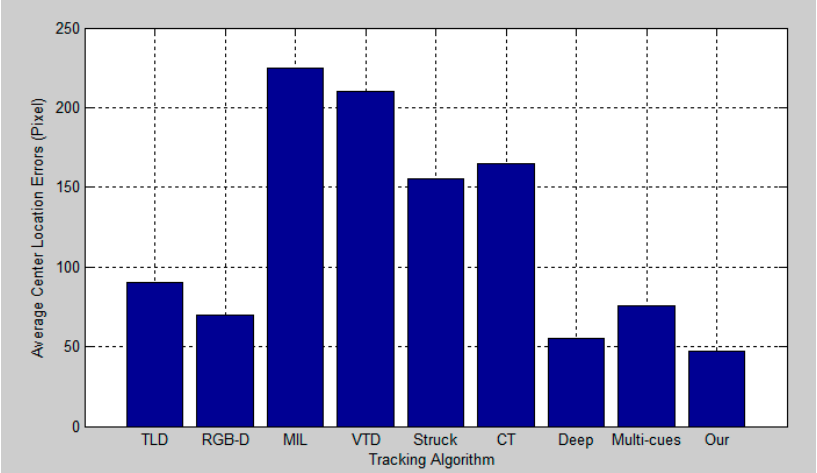

Figure 14. Quantitative evaluation in terms of average center location error (in pixel) for the third experiment.

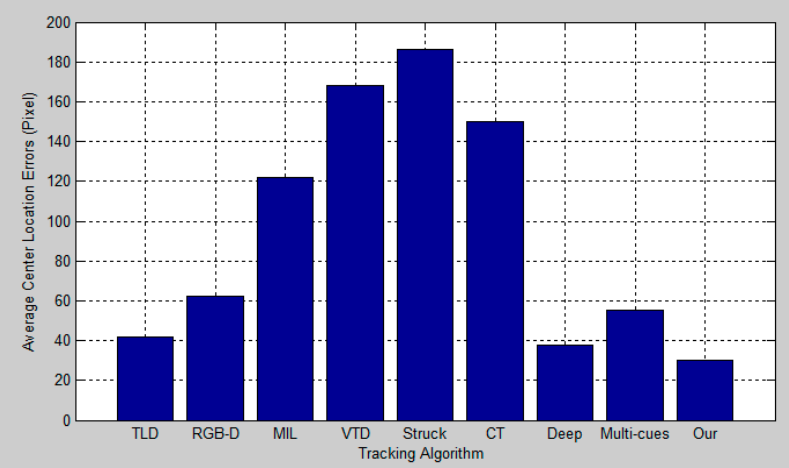

Figure 15. Quantitative evaluation in terms of average center location error (in pixel) for the fourth experiment.

Table 1 reports the success rates, where larger scores mean more accurate results.

Table 1. The evaluation results of SR under different categorizations.

Method	Object Type		Movement		Occlusion	
	Human	Animal	Fast	Slow	Yes	No
Our Tracker	80.1%	72.9%	77.5%	82.3%	81.2%	82.6%
TLD Tracker	29.0%	35.1%	29.7%	51.6%	33.8%	38.7%
VTD Tracker	30.9%	48.8%	37.2%	57.3%	28.3%	63.1%
MIL Tracker	32.2%	37.2%	31.5%	45.5%	25.6%	49.0%
RGB-D Tracker	47.1%	47.0%	51.8%	56.7%	46.9%	61.9%
Struck Tracker	35.4%	47.0%	39.0%	58.0%	30.4%	63.5%
CT Tracker	31.1%	46.7%	31.5%	48.6%	34.8%	46.8%
Deep Tracker	72.1%	64.8%	70.1%	76.3%	71.4%	72.6%
Multi-cues Tracker	33.2%	49.5%	52.3%	55.6%	44.7%	57.5%

Table 2 lists the average speed of each method on the recent benchmark dataset [21]. The average speed of our method is 0.14 fps, implemented in Matlab without optimization for speed. The fine-tuning of our method is time-consuming.

Table 2. The average speed of each method on the recent benchmark dataset [21].

Method	The Average Speed (fps)
Our Tracker	0.14
TLD Tracker	28.5
VTD Tracker	6.7
MIL Tracker	38.9
RGB-D Tracker	2.6
Struck Tracker	20.8
CT Tracker	64.7
Deep Tracker	0.23
Multi-cues Tracker	40.7

6. Conclusions

By analyzing the problems of the existing technologies, this paper proposes a visual object tracking algorithm based on cross-modality features learning using Gaussian-Bernoulli deep Boltzmann machines (DBM) over RGB-D data. We extract cross-modality features of the samples in RGB-D video data based on across-modality Gaussian-Bernoulli DBM and obtain the object tracking results over RGB-D data using aBayesian maximum a posteriori probability estimation algorithm. The experimental results show that the proposed method greatly improves the robustness and accuracy of thealgorithm. In the future, we will extend the proposed method to solve other vision problems (e.g., object detection, face recognition, etc.).

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China under Grant Nos.: 61403060, 61332017, 61602202, 61402192 and 61603146, in part by Dalian Science and Technology Planning Project under Grant 2015A11GX021, and the six talent peaks project in Jiangsu Province under Grant XYDXXJS-012 and XYDXXJS-011, in part by National Science and technology support project under Grant 2015BAK04B05, in part by National Natural Science Foundation of Jiangsu Province under Grant BK20160427, BK20160428.

Author Contributions: Mingxin Jiang and Zhigeng Pan conceived and designed the experiments; Zhenzhou Tang performed the experiments; Mingxin Jiang wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Coppi, D.; Calderara, S.; Cucchiara, R. Transductive People Tracking in Unconstrained Surveillance. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *26*, 762–775. [CrossRef]
2. Daniel, P.J.; Doherty, J.F. Track Detection of Low Observable Targets Using a Motion Model. *IEEE Access* **2015**, *3*, 1408–1415.
3. Doulamis, A. Dynamic tracking re-adjustment: A method for automatic tracking recovery in complex visual environments. *Multimed. Tools Appl.* **2010**, *50*, 49–73. [CrossRef]
4. Wang, B.X.; Tang, L.B.; Yang, J.L.; Zhao, B.J.; Wang, S.G. Visual tracking based on extreme learning machine and sparse representation. *Sensors* **2015**, *15*, 26877–26905. [CrossRef] [PubMed]
5. Li, X.L.; Han, Z.F.; Wang, L.J.; Lu, H.C. Visual Tracking via Random Walks on Graph Model. *IEEE Trans. Cybern.* **2016**, *46*, 2144–2155. [CrossRef] [PubMed]
6. Munaro, M.; Basso, F.; Menegatti, E. Tracking People within Groups with RGB-D Data. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems(IROS), Vilamoura, Portugal, 7–11 October 2012.
7. Spinello, L.; Luber, M.; Arras, K.O. Tracking people in 3D using a bottom-up top-down people detector. In Proceedings of the International Conference on Robotics and Automation (ICRA), Shanghai, China, 9–13 May 2011.
8. Spinello, L.; Arras, K.O. People Detection in RGB-D Data. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems(IROS), San Francisco, CA, USA, 25–30 September 2011.
9. Viola, P.; Jones, M.J. Robust real-time face detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
10. Navneet, D.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005.
11. Wang, X.Y.; Han, T.X.; Yan, S.C. An HOG-LBP human detector with partial occlusion handling. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan, 29 September–2 October 2009.
12. Li, H.X.; Li, Y.; Porikli, F. DeepTrack: Learning Discriminative Feature Representations Online for Robust Visual Tracking. *IEEE Trans. Image Process.* **2016**, *25*, 1834–1848. [CrossRef] [PubMed]
13. Felzenszwalb, P.; Girshick, R.; McAllester, D.; Ramanan, D. Object Detection with Discriminatively Trained Part Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [CrossRef] [PubMed]
14. Lu, J.W.; Wang, G.; Deng, W.H.; Moulin, P.; Zhou, J. Multi-manifold deep metric learning for image set classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
15. Wang, N.Y.; Yeung, D.Y. Learning a deep compact image representation for visual tracking. In Proceedings of the Conference and Workshop on Neural Information Processing Systems (NIPS), South Lake Tahoe, NV, USA, 5–10 December 2013.
16. Wang, L.; Liu, T.; Wang, G.; Chan, K.L.; Yang, Q.X. Video Tracking Using Learned Hierarchical Features. *IEEE Trans. Image Process.* **2015**, *24*, 1424–1435. [CrossRef] [PubMed]
17. Ding, J.; Huang, Y.; Liu, W.; Huang, K.Q. Severely Blurred Object Tracking by Learning Deep Image Representations. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *26*, 319–331. [CrossRef]
18. Kalal, Z.; Mikolajczyk, K.; Matas, J. Tracking-learning-detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1409–1422. [CrossRef] [PubMed]
19. Babenko, B.; Yang, M.H.; Belongie, S. Visual tracking with online multiple instance learning. In Proceedings of the IEEE International Conference on Computer Vision (CVPR), Miami Beach, FL, USA, 20–21 June 2009.
20. Kwon, J.; Lee, K.M. Visual tracking decomposition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010.
21. Song, S.; Xiao, J.X. Tracking revisited using RGBD camera: Unified benchmark and baselines. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Darling Harbour, Sydney, 3–6 December 2013.
22. Hinton, G.E.; Sejnowski, T.J. Optimal perceptual inference. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 8–10 June 1983.

23. Keronen, S.; Cho, K.; Raiko, T.; Ilin, A.; Palomäki, K.J. Gaussian-Bernoulli restricted Boltzmann machines and automatic feature extraction for noise robust missing data mask estimation. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP), Vancouver, BC, Canada, 26–30 May 2013.
24. Salakhutdinov, R.; Hinton, G.E. Deep Boltzmann Machines. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), Clearwater Beach, FL, USA, 16–18 April 2009.
25. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal Deep Learning. In Proceedings of the International Conference on Machine Learning (ICML), Bellevue, WA, USA, 28 June–2 July 2011.
26. Srivastava, N.; Salakhutdinov, R. Multimodal Learning with Deep Boltzmann Machines. In Proceedings of the International Conference and Workshop on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 3–8 December 2012.
27. Jiang, M.X.; Li, M.; Wang, H.Y. Visual Object Tracking Based on 2DPCA and ML. *Math. Probl. Eng.* **2013**, *2013*, 404978–404985. [CrossRef]
28. Luber, M.; Spinello, L.; Arras, K.O. People tracking in RGB-D Data with on-line boosted target models. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), San Francisco, CA, USA, 25–28 September 2011.
29. Hare, S.; Saffari, A.; Torr, P.H.S. Struck: Structured output tracking with kernels. In'Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 263–270.
30. Zhang, K.; Zhang, L.; Yang, M.-H. Real-Time Compressive Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Firenze, Italy, 7–13 October 2012; pp. 864–877.
31. Ruan, Y.; Wei, Z. Real-Time Visual Tracking through Fusion Features. *Sensors* **2016**, *16*, 949. [CrossRef] [PubMed]
32. Kuen, J.; Lim, K.M.; Lee, C.P. Self-taught learning of a deep invariant representation for visual tracking via temporal slowness principle. *Pattern Recognit.* **2015**, *48*, 2964–2982. [CrossRef]