ROBUST OBJECT TRACKING METHOD BY USING 3D SPATIAL-TEMPORAL MARKOV RANDOM FIELD

Huijun He, Hong Liu

Key Laboratory on Machine Perception and Intelligence, Peking University, P.R.China
Email:{hehj, hongliu}@cis.pku.edu.cn

Keywords: Tracking; Markov Random Field; Energy Function; Information Integration.

Abstract

In this paper a novel tracking method based on 3D Spatial-Temporal Markov Random Field (3D S-T MRF) is proposed. By taking temporal axis into account the traditional spatial MRF is extended to 3-D. Object tracking can be regarded as a labeling problem, i.e. assigning every pixel a label 0 or 1, of which 1 stands for tracking region and vice versa. Through defining proper 3D MRF structure, such as the nodes, neighbor system, data energy and smooth energy function, color cue and motion cue can be fused naturally. The labeling problem comes down to an energy minimization problem. Considering the simplicity and efficiency, Iterated Conditional Method (ICM) algorithm is used to minimize the energy function. The experiments show that this method can get promising results in challenging background and to some extent is robust against occlusions owing to the fusion of motion and color information through energy function.

1 Introduction

An efficient objects tracking algorithm in complex environments is a challenging task. Recent years much progress has been made in this domain and various algorithms have been proposed. Generally speaking, two major kinds of methods can be concluded, i.e. deterministic and probabilistic approaches. As the deterministic approach, Comaniciu et al. [2] proposed the mean-shift method which estimates the non-parametric density gradient based on color histogram. This method is quite effective and can handle partial occlusions. But it is hard to continue if complete occlusion occurs. As the probabilistic approach, Isard et al. [5] proposed CONDENSATION algorithm, also called particle filtering or bootstrap filtering. As multi-modal feature based tracking, Liu et al. [4] presents a Bayesian network based multi-modal fusion method for robust and real-time face tracking.

At present, object tracking under occlusions or clutter environments is still a challenging problem. In Mean Shift algorithm, if the object is totally occluded in two consecutive frames, the algorithm has difficulties in keep tracking. Moreover, because it presents the object using only kernel estimate of color distribution, when in clutter environments or there are similar objects around, it will be influenced greatly. A common way to handle occlusion is to fuse the information of object motion under Mean Shift framework.

Markov Random Field model in computer vision applications was first interpreted and used by Geman[8]. After that, MRF model has been employed successfully and widely in image applications such as restoration and segmentation [3,6]. In this paper the traditional MRF is extended to 3D Spatial-Temporal by taking time axes into account, a similar idea was proposed at [9]. Shunsuke KAMIJO used this model to track vehicle by referring to local motion vectors, texture and labelling correlations. However they don’t use the color information and the model is only suitable for rigid object such as vehicle. In our paper the color and motion cues are fused through energy function and can be used for non-rigid object tracking such as human body.

The rest of this paper is organized as follows. In section 2 we give a precise description of the 3D MRF model and derive the energy minimization problem. Section 3 shows the experimental results of our proposed method. Conclusions are made in section 4.

2 Tracking Algorithm

2.1 3D Spatial-Temporal MRF Model

As well known, markov chain is defined in time domain to process sequential data such as speech signal. Take the Hidden Markov Model (HMM) for example, this model consists of a finite set of states, each of which is associated with a probability distribution. Generally speaking, it can be called 1-D MRF.

As to 2D MRF which is most common in image applications, it’s defined in space domain, namely, *(i, j)* lattice of an image. Given observation O and some constraints, it’s required to obtain scene S to make the posterior probability maximum.

$$s^* = \arg\max P(S = s \mid O = o) \qquad (1)$$

According to Bayesian and Hammersley-C1ifford theorem the follow equation can be derived:

$$\begin{aligned} s^* &= \arg\max P(S = s \mid O = o) \\ &\propto \arg\max P(S = s)P(O = o \mid S = s) \\ &\propto \arg\min\{U(s) + U(o \mid s)\} \end{aligned} \tag{2}$$

Usually, *U(s)* is called smooth energy and *U(o|s)* data energy. Then solving the MRF model will come down to an energy minimization problem.

Note that video is a series of image sequence, by adding time axes 2D MRF is extended naturally to 3D MRF defined in *(x, y, t)* space. Consequently tracking one object in an image sequences is equivalent to inferring the label of each pixel according to its observation, i.e. inferring label field *L* from observation field *O*.

In current frame, the coordinates of each pixel *p* can be written as *p = (i, j, t)*, which associates a corresponding hidden state node *s(i, j, t)* or called the label and the data node *d(i, j, t)* i.e. the observation. Each pixel interacts with in spatial domain *(i±1, j, t)* and *(i, j±1, t)*, and in time domain *(i, j, t-1)* which is the counterpart in the previous frame *t-1*. In sum, each pixel has four spatial neighbors and one temporal neighbor and the decision of its label receives influence from the five neighbors through potential functions (also called energy function). The total energy is the sum of the two following terms:

$$E(l,o) = E(l) + \lambda E(o,l) \tag{3}$$

The first term denotes the smooth energy representing the prior constraints between neighboring pixels, while E(o,l) denotes data energy representing the likelihood of the label given observation. λ is the constant to control the weight of data energy.

2.2 Definition of Energy Functions

The energy function includes two terms, data energy and smooth energy. The smooth energy is define as

$$E(l) = \sum_{c \in C} V_c(l_i, l_j) \tag{4}$$

Here l_i and l_j are the neighbouring pixel in a clique, and $V_c(l_i,l_j)$ is the smooth energy associated with every clique *c*. In order to impose the smooth constrains to the neighbor and remove the isolated points caused by noise, the energy function is defined as Potts model,

$$V_c(l_i, l_j) = \begin{cases} -\beta, l_i = l_j \\ \beta, l_i \neq l_j \end{cases} \tag{5}$$

This term is the smooth constraint which makes the labels of neighboring pixels tend to be the same. The parameter β can be different for spatial neighbor and temporal neighbors.

The data energy is defined as follows,

$$E(o,l) = \ln \prod_{i \in I} \frac{1}{p_i(o \mid l)} = \sum_{i \in I} - \ln p_i(o \mid l) \tag{6}$$

For each pixel, *p_i(o|l)* represents the probability that observation o belongs to tracking region and $lnp_i(o|l)$ is the log likelihood. To integrate motion cue and color cue, for pixel *p(i,j)* its likelihood *p(o|l)* is computed as follows.

Assume that the background and the camera is static, which is common in really world applications. At first the probability density function (*PDF*) of the background and the target region are estimated. Owing to the static background and aims of reducing the computing complexity, the estimate of background probability density is only computed once using kernel method and then stored. Usually the kernel function is the Guass kernel:

$$G_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{\sigma^2}} \tag{7}$$

Given the sample set $S=\{x_i\}$ taking from the image the density estimate at any point y can be evaluated as follows[1],

$$\tilde{p}_0(y) = \frac{1}{N\prod_{j=1}^{d}\sigma_j} \sum_{i=1}^{N} \prod_{j=1}^{d} G(\frac{y_j - x_{ij}}{\sigma_j}) \tag{8}$$

,

where the same kernel is used in each dimension with different bandwidth σ_j. Usually a weighted version can be used in which bigger weight is given to samples inside the region and smaller weight for samples from the boundary.

For the target region $\tilde{p}_1(y)$, it should be evaluated each frame because the object may be non-rigid and the appearance changes over time. Therefore the simple histogram method is adopted instead of the expensive kernel estimate. The initial target region is selected manually as a rectangle in initial step.

For each pixel, back project the color observation which is represented by vector $y_{i,j}(H,S,V)$ to the background density $\tilde{p}_0(y)$ and the object density $\tilde{p}_1(y)$, and hence the color likelihood p_{color_0} and p_{color_1} can be expressed as $\tilde{p}_0(y_{i,j})$ and $\tilde{p}_0(y_{i,j})$ respectively.

Background subtraction method is adopted as the motion cue to detect the rough body area on which the accurate area is estimated. For the aim of computing simply, single Guass model is used. The gray level of each pixel is represented by Guassian distribution $N(y;\mu,\sigma^2)$. The parameter μ and σ is estimated according to the first several frames in which the tracked foreground is absent. With the time going, the background model is updated.

The probability that one pixel belongs to background is

$$p_{motion\_0} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_0 - \mu)}{2\sigma^2}} \tag{9}$$

Equivalently, the probability of belonging to the tracking object is

$$p_{motion\_1} = 1 - \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y_0-\mu)}{2\sigma^2}} \qquad (10)$$

It is defined as the motion likelihood.

The minimum method is used to integrate the two cues:

$$p(o \mid l) = \begin{cases} \min(p_{color\_1}, p_{motion\_1}), l = 1 \\ \min(p_{color\_0}, p_{motion\_0}), l = 0 \end{cases} \qquad (11)$$

Hence the log likelihood and the data energy can be obtained.

2.3 Improved Neighbour Definition in Occlusion

To resolve occlusion necessarily need to model the object motion because the observation data is incomplete in these circumstances. Look back the definition of neighbor system, it can be found that the temporal neighbor is not suitable in occlusion situation. Therefore, it's reasonable to make use of the estimated motion vector to redefine the temporal neighbor. Assume that in frame t occlusion happens, the motion vector V=($v_{x,}$ v_y) can be estimated according to the locations tracked in former N frames. The temporal neighbor of pixel (i, j) is the one shifted backward in the amount of the motion vector V=($v_{x,}$ v_y) in previous frame.

The algorithm will judge whether the tracked target is occluded by computing the number of labels assigned to the target which is denoted as *N*. Set the threshold *Th* in advance and then compare *N* with it. If in one frame the number of labels satisfies *N-Th<0*, the color distribution of the target $\tilde{p}(y)$ in that frame will be kept and the new neighbor definition based on motion vector works, which is described in Fig. 1:

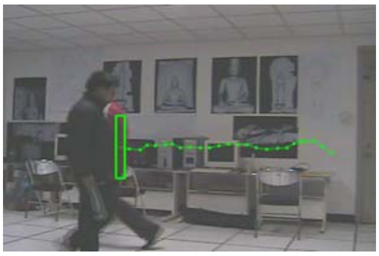


Fig.1. The new definition of temporal neighbor works

when occlusion is detected

2.4 ICM Solution

Solving the MRF model comes down to an optimization problem. Iterated Conditional Modes (ICM) [7] uses a deterministic greedy strategy to find a local minimum. It initializes with an estimate of the labeling, and then for each pixel, selects the label whose energy function is the lowest. This process is repeated until convergence. However, the ICM method is extremely sensitive to the initial estimate. Generally it is initialized by assigning each pixel the label with the lowest data cost i.e. the larger likelihood probability and it is proper demonstrated by the result. Generally the diagram of this method is depicted as Fig.2.

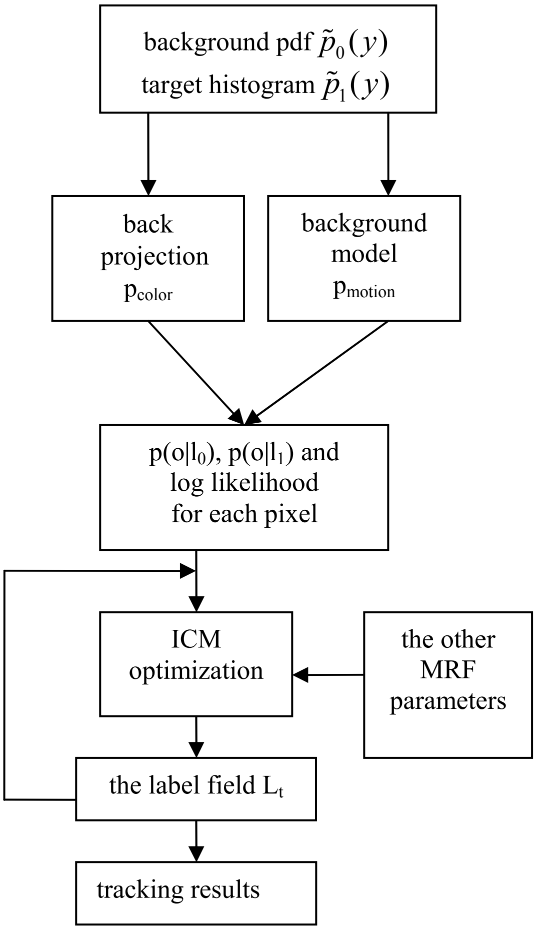
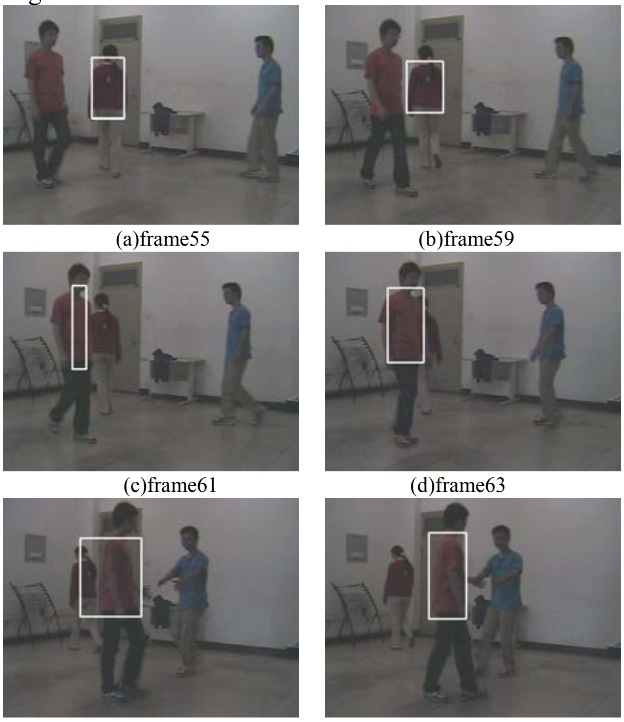

background pdf $\tilde{p}_0(y)$
target histogram $\tilde{p}_1(y)$
back projection p_{color}
background model p_{motion}
p(o|l$_0$), p(o|l$_1$) and log likelihood for each pixel
ICM optimization
the other MRF parameters
the label field L$_t$
tracking results

Fig.2. the diagram of MRF based tracking algorithm.

3 Experimental results

This method is implemented on 6 video sequences captured at the rate of 12 frames/second in an indoor laboratory environment. HSV color space is used considering that RGB space is sensitive to illuminative variations. When pixels' color has a small saturation near zero, hue is sensitive and inaccurate, which results in inaccuracy and noise in the back projection image. In view of this, low limits are set for saturation and value.

Totally, objects in 86% frames can be tracked successfully. As to the reasons for failure, one is that the color distributions of background and objects are too similar, another is that the object may go out of the view and the algorithm will fail. The performances on one typical video are shown in Fig.3 and Fig.4 compared with mean-shift method. The tracked object is the girl in red coat. At the same time, one boy in similar color comes near. As demonstrated in Fig.3, the result rectangle tends to transfer and fails because mean shift algorithm only takes into account the color information. Fig4 is the results produced by 3D MRF algorithm, which considers not only the appearance and color cues but also the motion and spatial information through energy function. Consequently, this algorithm is robust to occlusion to some extent.

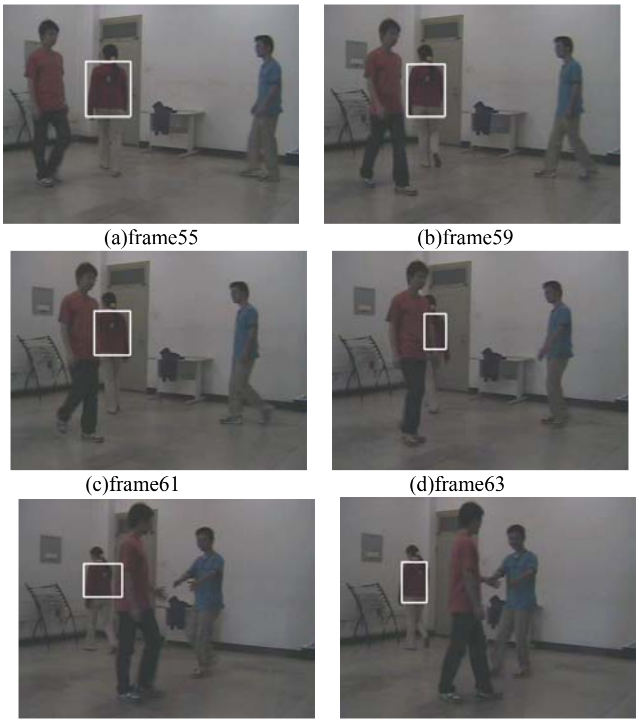


(a)frame55 (b)frame59

(c)frame61 (d)frame63

(e)frame71 (f)frame73

Fig.3.results of mean shift algorithm



(a)frame55 (b)frame59

(c)frame61 (d)frame63

(e)frame71 (f)frame73

Fig.4. results of the algorithm proposed in this paper

There are yet several aspects that can be improved and advanced. For example, given that the tracker is applied to human, we can put the prior shape and edge constraints on the data energy and can expect better performance.

4 Conclusions

In this paper, a 3D spatial-temporal MRF method is proposed for visual tracking. Tracking can be expressed elegantly as a labeling problem under this method. The total energy is the function of the label field and solving the labels comes down to an energy minimization problem. 3D MRF model can integrate appearance and motion cues naturally. What's more, it considers not only the spatial constraints through proper smooth energy but also the temporal constraints imposed by the neighbor frame through data energy. The experimental results show that this method can obtain robust performance in occlusion situation.

Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC, No. 60675025), National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247) and the Project of Shenzhen Science and Technology in 2008.

References

[1] A. Elgammal, R. Duraiswami, L.S. Davis: Efficient kernel density estimation using the fast gauss transform with applications to color modeling and tracking. IEEE Trans. PAMI, Vol.25, Issue.6, pp1499-1504, 2003.

[2] D. Comaniciu, P. Meer: Real-Time Tracking of Non-Rigid Objects Using Mean Shift. IEEE Conf. CVPR, Vol.2, pp142-149, 2000.

[3] D. Geman and G. Reynolds: Constrained Restoration and the Recovery of Discontinuities, IEEE Trans. PAMI, Vol.14, No.3, pp367-383, 1992.

[4] F. Liu, X. Lin, S. Li and Y. Shi: Multi-Modal Face Tracking Using Bayesian Network, IEEE Workshop, AMFG, pp135-142, 2003.

[5] M. Isard, A. Blake: CONDENSATION – Conditional Density Propagation for Visual Tracking, IJCV1 (29), pp5-28, 1998.

[6] P. Andrey, P. Tarroux: Unsupervised Segmentation of Markov Random Field Modeled Textured Images Using Selectionist Relaxation. IEEE Trans. PAMI, Vol.20, No.3, pp252-262, 1998.

[7] R. Szeliski and R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, C. Rother: A Comparative Study of Energy Minimization Methods for Markov Random Fields. ECCV, Vol. II, pp16-29, 2006.

[8] S. Geman and D. Geman: Stochastic Relaxation, Gibbs Distribution, and the Bayesia Restoration of images, IEEE Trans. PAMI, Vol.6, No.6, pp721-741, 1984.

[9] S. Kamijo: Spatio-Temporal MRF model and its Application to Traffic Flow Analyses. Proceedings of the 21st International Conference on Data Engineering Workshops, pp1203-1211, 2005.