ADAPTIVE COOPERATIVE TRACKING BASED ON MULTI-GRAPH EMBEDDING AND MARKOV RANDOM FIELD

Lin Ma[1], *Junliang Xing*[1], *Xiaoqin Zhang*[2], *Weiming Hu*[1]

1. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
2. College of Mathematics & Information Science, Wenzhou University, China
E-mail: lin.ma@ia.ac.cn

ABSTRACT

Appearance model is of fundamental importance in a tracking algorithm. In this paper, we propose a new tracking method based on a cooperative object appearance model which incorporates both the discriminative and generative information. We represent the discriminative information with graph embedding (GE). To represent the local object appearance effectively, we divide the object and nearby background into patches. As the discriminative conditions around the 4 object boundaries are different, we divide the patches into 4 groups and perform GE for each group. Markov Random Filed (MRF) is designed to represent the generative information. We propose a novel MRF based method which not only considers the single patch's appearance but also the appearance relations between neighbor patches (not the relations between neighbor patches' states). The proposed cooperative appearance model can represent the object appearance's variation effectively and meanwhile discriminate the object from background robustly. Experimental results on challenging test sequences demonstrated the effectiveness of our method.

Index Terms— tracking, graph embedding, MRF

1. INTRODUCTION

Object tracking is a popular research field in computer vision. To obtain promising tracking results, various appearance models [1, 2, 3, 4] are adopted for object tracking. The appearance models can be roughly categorized into two groups: generative and discriminative. Generative model represents the foreground information effectively by representing the samples' distributions etc. Bradski [5] proposes a color histogram-based method: Camshift (Continuously Adaptive Mean-Shift), to track faces by utilizing the color statistics information. Ross et al. [6] use incremental PCA (Principal Components Analysis) to represent object appearances' distributions and obtain promising tracking results. Generative model represents the foreground's variation effectively and is less easily influenced by drastic background variation than discriminative model. But when the background varies not drastically, the discriminative model [7, 8] is generally more robust and able to tackle the drift problem etc. more effectively. Babenko et al. [9] perform tracking with online Multiple Instance Learning (MIL). MIL remains both positive and negative samples and the discriminative information between the two kinds of samples is utilized to determine the object state. GE is an important learning method which unifies different dimensionality reduction methods [10, 11]. Ma et al. [12] form multiple sample group pairs to obtain accurate discrimination of GE for tracking. Kernel method is also utilized in GE to solve nonlinear problems by some researchers [13].

Our work is mostly related to [12], which also divides object appearance and nearby background into patches and performs GE for multiple sample groups. However, in [12] the groups are formed based on classification of foreground patches, while our method forms 4 groups around 4 object boundaries respectively and performs GE for each group separately. Compared with [12], the new method represents the local discrimination more accurately. Moreover, an MRF based generative model is also introduced in our method. Yang and Wu [14] construct the interest region based graph and perform tracking based on MRF method. However, they only consider the single node's appearance, and do not consider the appearance relations between different nodes. In contrast, we consider both the two appearances in MRF and obtain more effective representation of object appearances.

In this paper, the GE and MRF information are combined together in the novel appearance model. The proposed cooperative appearance model is deployed into the Bayesian inference framework with a particle filter implementation to perform tracking. By combining the generative information and the discriminative information, the proposed appearance model can represent the object appearance's variation effectively and meanwhile discriminate the object from background robustly. The flowchart of our method is shown in Fig. 1, and the key contributions of our paper are as follows.

(1) We perform GE on 4 patch groups. In this way, we are able to compute the local discriminative information more accurately.
(2) We propose a novel MRF method which represents two kinds of patch appearances effectively.
(3) We combine together GE and MRF, which are discriminative model and generative model respectively, to perform tracking.

The rest of our paper is organized as follows: Section 2 shows the particle filter we use in this paper. In Section 3, we evaluate the particles with the multi-graph embedding. In

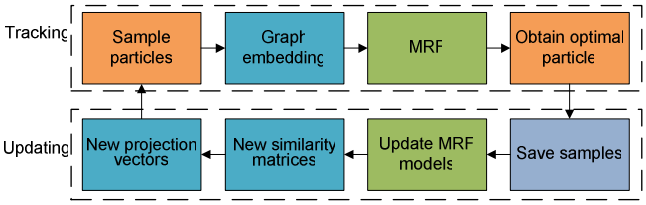


Fig. 1. Flowchart of our method. We perform tracking in the framework of particle filter. The particles are evaluated with GE and MRF.

Section 4, we propose a novel MRF method to evaluate the particles. Experimental results are presented in Section 5, and conclusion and future work are shown in Section 6.

2. PARTICLE FILTER FRAMEWORK

Particle filter represents the distributions of object's sates effectively and is widely used in tracking [15, 16]. In this paper, we perform tracking using the Bayesian inference method with a particle filter. Given observation sequence $O_{1:t+1}$, the posterior probability density of the object state is defined as

$$p(X_{t+1}|O_{1:t+1}) \propto p(O_{t+1}|X_{t+1}) \int p(X_{t+1}|X_t) p(X_t|O_{1:t}) dX_t. \quad (1)$$

where X_t represents the object state at time t. We use the same definition of X_t and the same definition of state translation as in [12]. Each particle is warped to a normalized 32×32 sub image. Detailed presentation of particle filter can be found in [15]. We select the particle with the largest likelihood $\pi_t = p(O_t|X_t)$ as the optimal particle. The particles are evaluated with GE and MRF.

3. MULTIPLE GROUPS' GRAPH EMBEDDING

3.1. Graph embedding

GE is a framework of dimensionality reduction and can unify most dimensionality reduction algorithms (such as PCA, LDA, etc.) in the same framework [10, 11]. Let $x_i \in R^D, i = 0, \ldots, N-1$ be D-dimension samples (here D=64), and $y_i \in \{0, \ldots, C-1\}$ be x_i's class label. n_c is defined as the number of samples belonging to class c, and satisfies $\sum_{c=0}^{C-1} n_c = N$. To represent the local information of the object appearance accurately, we divide the object appearance and nearby background to 6×6 patches (Fig. 2(a)). Each patch is a sample. Let W be a similarity matrix, the element $w^*_{i,j}$ of which represents the similarity between samples x_i and x_j. With the sample set $\mathcal{X} = \{x_0, \ldots, x_{N-1}\}$ and W, we construct an undirected graph $G = \{\mathcal{X}, W\}$. Let the element $d_{i,i}$ of the diagonal matrix $\widetilde{D}$ be

$$d_{i,i} = \sum\nolimits_{j \neq i} w^*_{i,j}, \quad (2)$$

and Laplacian matrix L be

$$L = \widetilde{D} - W. \quad (3)$$

Then the projection vectors is obtained by solving

$$P^* = \arg\min_{P^T Z^T B Z = I} \sum_{i,j} \|z_i - z_j\|^2 w^*_{i,j} = \arg\min_{Z^T B Z = I} 2\,\mathrm{tr}(Z^T L Z), \quad (4)$$

where $z_i = P^T x_i$, tr(v) represents the trace of the matrix v. With P^*, we are able to obtain the low dimensional representations of the samples.

We perform GE with mean foreground samples and mean background samples. We define $\overline{I}^f_{i,t}$ as the mean of foreground sample i of frame $0,\ldots,t$. As normally the background varies largely, we define background sample j's mean $\overline{I}^b_{j,t}$ with only recent 5 frames, and when define samples' variances in Section 3.2 we perform in the same way. The near samples (patches) often have larger similarity, and vice verse. To discriminate the foreground and background samples more accurately, we divide the 6×6 samples into 4 groups. Each group's samples are around one of the 4 object boundaries and formed of 8 foreground samples and 10 background samples. We perform GE for each group separately. For group $k = 0,..,3$, let $\overline{u}^{(k)}_f$ be the mean of the foreground samples of all frames, and let $\overline{u}^{(k)}_b$ be the mean of the background samples of the recent 5 frames. Then similar to [12], we define $u^{(k)} = (\overline{u}^{(k)}_f + \overline{u}^{(k)}_b)/2$ as the mean of all samples of group k. The sample vector is reduced to 1D subspace in our paper.

Let $\mathcal{X}^{(k)}$ be a matrix whose columns are the foreground and background samples of group k and e be an all 1 vector. We define $\mathcal{X}^{(k)}_0 = (\mathcal{X}^{(k)} - u^{(k)} e^T) B^{(k)}$, where $B^{(k)} = \mathrm{Diag}(b_{k,0}, ..., b_{k,17})$ defines the coefficient of each column of $\mathcal{X}^{(k)} - u^{(k)} e^T$. Then the projection vector p_k is obtained by optimizing the objective function

$$J(p_k) = p_k^T \mathcal{X}^{(k)}_0 L \mathcal{X}^{(k)T}_0 p_k. \quad (5)$$

$$s.t. \quad p_k^T p_k = 1$$

With Lagrange method, the optimal p_k corresponds to the largest λ_k in

$$\mathcal{X}^{(k)}_0 L \mathcal{X}^{(k)T}_0 p_k = \lambda_k p_k. \quad (6)$$

In [12], samples' coefficients, i.e. $B^{(k)}$, are revised according to the distances between samples and the discriminative plane. Here, we use the same method to define $B^{(k)}$. The definition of the similarity matrix W_k is detailed in Section 3.2. Let the mean of foreground patches of group k in frame t be $u^{(k)}_{f,t}$. $u^{(k)}_{f,t}$ normally is near to $\overline{u}^{(k)}_f$ and far from $\overline{u}^{(k)}_b$. Let

$$E_k = \left| p_k^T \left(u^{(k)}_{f,t} - \overline{u}^{(k)}_b \right) / D \right| - \left| p_k^T \left(u^{(k)}_{f,t} - \overline{u}^{(k)}_f \right) / D \right|. \quad (7)$$

Then the likelihood corresponding to GE is defined as

$$p(O^{ge}_t | X_t) = \exp\left\{ \sum\nolimits_{k=0}^{3} w_{(k)} E_k \right\}, \quad (8)$$

where $w_{(k)}$ is the weight of group k. The proposed multiple groups' GE method is summarized in Algorithm 1.

3.2. Inter-class similarity

We define the similarity matrix based on samples' steadiness. Let the variance of the foreground sample i=0,...,15 be $\sigma^2_{f,i}$, and background sample j=0,...,19 be $\sigma^2_{b,j}$. We define $\overline{\sigma}^2_f =$

$\frac{1}{16}\sum_i \sigma_{f,i}^2$ and $\overline{\sigma}_b^2 = \frac{1}{20}\sum_j \sigma_{b,j}^2$. For group k, we define the similarity matrix $W_k = \begin{bmatrix} 0 & W_k' \\ W_k'^T & 0 \end{bmatrix}$, where W_k' is the similarity matrix of samples belonging to foreground and background respectively. Normally we consider that more steady samples (of smaller variances) are more confident in computing p_k. Let $\alpha_{k,i}$ and $\beta_{k,j}$ be the confidences of foreground sample i and background sample j of group k respectively. Then we define

$$\alpha_{k,i} = \exp\left(-\sigma_{f,i'}^2/\overline{\sigma}_f^2\right), \quad (9)$$

$$\beta_{k,j} = \exp\left(-\sigma_{b,j'}^2/\overline{\sigma}_b^2\right), \quad (10)$$

and define

$$w_{i,j}^{\prime(k)} = \alpha_{k,i}\beta_{k,j}, \quad (11)$$

where $w_{i,j}^{\prime(k)}$ is a component of W_k' , i' and j' are corresponding patch indices of group k's samples i and j.

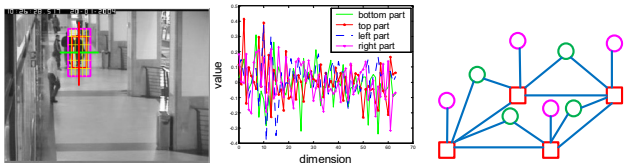
Algorithm 1 The proposed multiple groups' GE method

Input:$\overline{I}_{i,t}^f$, $\sigma_{f,i}^2$,$i = 0, \ldots, 15$; $\overline{I}_{j,t}^b$, $\sigma_{b,j}^2$, $i = 0, \ldots, 19$.
Steps: Form 4 groups of samples with the 36 samples. For group k, do:
1. Compute W_k according to (11).
2. Compute p_k.
1) Initialize $B^{(k)}$ with $b_{k,i} = 1$, $i = 0, \ldots, 17$.
2) Obtain p_k according to (6).
3) Revise $B^{(k)}$ with the method in [12] and do 2) of Step 2 again, until reach the maximum iteration times (2 in this paper).
Output: $p_k, k = 0, ..., 3$.

4. MRF LIKELIHOOD FOR OBJECT PATCHES

The multiple groups' GE method is able to discriminate the foreground areas from the background areas effectively. In another part, the object appearance (the generative information) is also important to represent the object, especially when background varies drastically. In this paper, we use MRF to represent the generative information. As Fig.2 shows, the MRF tracking problem can be defined as computing the hidden state given the known observation node and the links in the graph. In particle filter, the candidate object states are represented by a set of particles. Then the problem is turned to compute the optimal state (particle) corresponding to the largest likelihood given the observation model. Traditional MRF only considers the appearance O_i^s about single state node i, however our method considers both O_i^s and the appearance O_i^b representing the relations between neighbor patches (with shared edge). Compared with traditional method, our method is able to represent more appearance information.

We adopt Gaussian model to represent the two kinds of object appearances in MRF. In frame t, for foreground patch I_i^f, we define $I_i^f \sim N(\overline{I}_{i,t}^f, \sigma_{f,i}^2)$. Let $d_{i,j}'$ be the Euclidean



(a) (b) (c)

Fig. 2. Object patches. (a) Grid. Red rectangle: object state. Pink rectangle: contain background and foreground. Yellow lines separate patches. The patches are divided into 4 groups. Each group contain 18 patches, which are below the green line, above the green line, left of the red line, right of the red line respectively. (b) Projection vectors of the 4 groups. (c) The proposed MRF model. Each foreground patch corresponds to one state node (red rectangle). Pink circle: appearance (observation) node about single state node. Green circle: appearance node about neighbor state nodes.

distance between (the textures of) patch i and patch i's neighbor j, and let $\overline{d}_{i,j}$ and $\theta_{i,j}^2$ be the mean and variance about $d_{i,j}'$ of all frames. We define $d_{i,j}' \sim N(\overline{d}_{i,j}, \theta_{i,j}^2)$. According to the MRF model, patch i's state node and the corresponding observation node form a clique. In addition, state nodes of patch i and patch j (patch i's neighbor) and the appearance node corresponding to the two nodes also form a clique. Given the object state (particle), the cliques are independent of each other. Let $p_s(I_i^f)$ and $p_b(d_{i,j}')$ be the probability densities of I_i^f and $d_{i,j}'$ respectively, and let w_i be patch i's weight. Then we obtain the likelihood

$$p_0(O_t^{mrf}|X_t) = p(O_t^s, O_t^b|X_t) = p(O_t^s|X_t)p(O_t^b|X_t). \quad (12)$$

We define $p(O_t^s|X_t) \propto \prod_{i=0}^{15}(p_s(I_i^f))^{w_i}$ and $p(O_t^b|X_t) \propto \prod_{i,j=0,j\in neig(i)}^{15}(p_b(d_{i,j}'))^{w_i w_j}$, where $neig(i)$ represents patch i's neighbors. To make the values obtained with (12) not too small, the MRF likelihood is defined as

$$p(O_t^{mrf}|X_t) = \log(p_0(O_t^{mrf}|X_t)). \quad (13)$$

The combined likelihood of X_t is defined as

$$p(O_t|X_t) \propto p(O_t^{ge}|X_t)p(O_t^{mrf}|X_t). \quad (14)$$

As $p(O_t^{mrf}|X_t)<0$, $p(O_t^{mrf}|X_t)$ is normalized to [0,1] in (14), i.e. define the minimum value of $p(O_t^{mrf}|X_t)$ among all the particles as 0, the maximum value as 1, all other particles' values of $p(O_t^{mrf}|X_t)$ are projected into [0, 1] linearly. $p(O_t^{ge}|X_t)$ is processed in the same way. We update the GE appearance model and the MRF appearance model every 5 saved samples. For foreground patch i, if $||I_i^f - \overline{I}_{i,t}^f||_2^2 < \alpha\sigma_{f,i}^2$ we consider patch i is not occluded, where α is a constant. If the number of not occluded patches ($0 \sim 16$) is larger than a threshold, we store the current frame for system updating, otherwise the current sample is dropped. Normally the occlusion conditions in successive frames are similar. Thus, given foreground patch $I_{i,t-1}$ in frame t-1, we define $w_i \propto p_s(I_{i,t-1})$ and $w_{(k)} \propto \sum_{i\in G(k)} w_i$ in frame t, where $G(k)$ represents group k's samples.

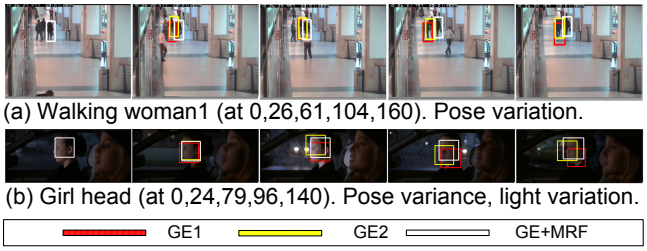


Fig. 3. Comparison between GE1, GE2 and GE+MRF.

Table 1. Average errors of the 3 methods in Fig. 3.

	GE1	GE2	GE+MRF
Fig. 3(a)	20.9	12.0	**4.2**
Fig. 3(b)	19.2	20.6	**8.0**

5. EXPERIMENTS

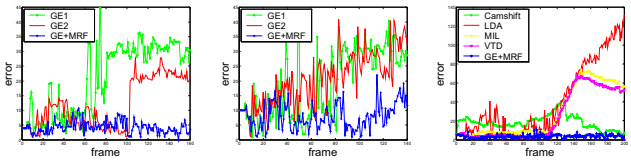
5.1. Experimental setting

We implemented our method in C++ and tested it on challenging videos. The experiments were conducted on a computer with an Intel 2.53 GHz CPU and 2G RAM. For each experiment of our method, we updated the system every 5 saved samples, and used 150 particles per frame during tracking. The object states in the first 5 frames were manually set. The running time of our method was around 0.1 sec/frame. We adopted the Euclidean distance between the object bounding box's center and the ground truth to represent the tracking error. Our method was performed on gray scale images. The test videos were from [17] and [18].

5.2. Experimental performance

In Fig. 3, we tracked a walking woman and the head of a girl, and compared our method (GE+MRF) with GE1 [12] and GE2. Here, we defined GE2 as only using the new GE method to evaluate particles. GE1, which considered the GE information and the histogram contrast between object and the nearby background, was able to discriminate the object from the background effectively. However, as in Fig. 3(a) when nearby background contained similar objects, GE1 was influenced largely and failed to track the object robustly any more. Similarly, GE2 also only used the discriminative information provided by GE and was not able to obtain accurate results. In contrast, the proposed MRF represented the local information of object appearance and relations between neighbor patches effectively. Thus, by combining GE and MRF, our method tracked the object accurately. Also, as in Fig. 3(b) when light varied largely, GE1 and GE2 were not able to discriminate the foreground from the background effectively. However, by taking advantages of both GE and MRF, our method was able to obtain more accurate results. The average errors and the error maps of the 3 methods were shown in Tab. 1 and Fig. 5 respectively.

Fig. 4. Walking woman2 (at 0,114,141,186,200). Occlusion. Comparison between our method and other 4 methods.



(a) walking woman1 (b) girl head (c) walking woman2

Fig. 5. Error maps.

In Fig. 4, we tracked a woman walking with another woman and compared our method with other 4 methods: Camshift [5], LDA, MIL [9] and VTD [3]. The object bounding box contained large background areas. As a result, the color histogram in Camshift was influenced by background, which made Camshift lose tracking. When the woman was occluded severely by a walking man, the object appearance varied largely. LDA and MIL were influenced by the man's appearance and not able to discriminate the foreground from the background effectively. Thus, the two methods failed to track the object robustly. The features adopted by VTD were also disturbed by the man's appearance and thus VTD also lost tracking. However, by giving different patches different weights according to the patches' distributions and by taking advantages of both discriminative and generative information, our method was able to tackle the occlusion problem effectively. The error map of Fig. 4 was shown in Fig. 5, and the average errors of Camshift, LDA, MIL, VTD and GE+MRF were 18.5, 40.4, 28.4, 24.4 and 3.5 respectively.

6. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a new method which combines GE and MRF to perform cooperative tracking. We divided the patches into 4 groups and performed GE for each group to obtain more accurate local discriminative information. We also proposed a new MRF method to obtain more effective object representation. The experiments demonstrated our method's effectiveness. In the future, we will continue the researches in combining the discriminative model and the generative model more effectively. (This work is partly supported by NSFC (Grant No. 60935002, 61100099, 61100147), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), and The Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081), NSF of Zhejiang Province (Grant No. LY12F03016).)

7. REFERENCES

[1] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation", *TPAMI*, 33(11), pp. 2259-2272, 2011.

[2] B. Liu, J. Huang, L. Yang, and C. Kulikowsk, "Robust tracking using local sparse appearance model and k-selection", *CVPR*, 2011.

[3] J. Kwon and K. M. Lee. "Visual Tracking Decomposition", *CVPR*, 2010.

[4] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan, "Locally orderless tracking", *CVPR*, 2012.

[5] G.R. Bradski, "Computer vision face tracking for use in a perceptual user interface", *Intel Tech. J.*, 2(Q2), 1998.

[6] D.A. Ross, J. Lim, R.S. Lin, and M.H. Yang, "Incremental learning for robust visual tracking", *IJCV*, 77(1), pp. 125-141, 2008.

[7] R. Hess and A. Fern, "Discriminatively trained particle filters for complex multi-object tracking", *CVPR*, 2009.

[8] G. Li, D. Liang, Q. Huang, S. Jiang, and W. Gao, "Object tracking using incremental 2D-LDA learning and Bayes inference", *ICIP*, 2008.

[9] B. Babenko, M. Yang, and S. Belongie, "Visual tracking with online multiple instance learning", *CVPR*, 2009.

[10] S. Yan, D. Xu, B. Zhang, H.J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction", *TPAMI*, 29(1), pp. 40-51, 2007.

[11] X.Q. Zhang, W.M. Hu, S. Maybank, and X. Li, "Graph based discriminative learning for robust and efficient object tracking", *ICCV*, 2007.

[12] L. Ma, W.M. Hu, and X.Q. Zhang, "Multiple sample group pairs' graph embedding for tracking", *ICIP*, 2012.

[13] P. Tiwari, J. Kuihanewicz, M. Rosen, and A. Madabhushi, "Semi supervised multi kernel (sesmik) graph embedding: identifying aggressive prostate cancer via magnetic resonance imaging and spectroscopy", *MICCAI*, 2010.

[14] M. Yang and Y. Wu, "Granularity and elasticity adaptation in visual tracking", *CVPR*, 2008.

[15] M. Isard and A. Blake, "CONDENSATION-conditional density propagation for visual tracking", *IJCV*, 1 (29), pp. 5-28, 1998.

[16] X.Q. Zhang, W.M. Hu, S. Maybank, and X. Li, "Sequential particle swarm optimization for visual tracking", *CVPR*, 2008.

[17] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

[18] L. Ivan, M. Marcin, S. Cordelia, and R. Benjamin, "Learning realistic human actions from movies", *CVPR*, 2008.