Visual Tracking Using Improved Multiple Instance Learning with Co-training Framework for Moving Robot

Zhiyu Zhou[1,*], Junjie Wang[1], Yaming Wang[1], Zefei Zhu[2], Jiayou Du[2], Xiangqi Liu[2], Jiaxin Quan[1]

[1] School of Information Science and Technology, Zhejiang Sci-Tech University
Hangzhou 310018, China
[email:13065713897@163.com]
[2] School of Mechanical Engineering, Hangzhou Dianzi University
Hangzhou, 310018, China
*Corresponding author: Zhiyu Zhou

Received August 8, 2017; revised June 18, 2017; accepted July 4, 2018; published November 30, 2018

Abstract

Object detection and tracking is the basic capability of mobile robots to achieve natural human–robot interaction. In this paper, an object tracking system of mobile robot is designed and validated using improved multiple instance learning algorithm. The improved multiple instance learning algorithm which prevents model drift significantly. Secondly, in order to improve the capability of classifiers, an active sample selection strategy is proposed by optimizing a bag Fisher information function instead of the bag likelihood function, which dynamically chooses most discriminative samples for classifier training. Furthermore, we integrate the co-training criterion into algorithm to update the appearance model accurately and avoid error accumulation. Finally, we evaluate our system on challenging sequences and an indoor environment in a laboratory. And the experiment results demonstrate that the proposed methods can stably and robustly track moving object.

***Keywords**:* object tracking, multiple instance learning, active learning, co-training, moving robot.

1. Introduction

Object tracking has been widely used in the various applications such as image compression and medical robot et al. [1, 2, 3, 4]. Tracking object [5,6] is a fundamental capability of mobile robot in a dynamic environment, understanding how people move in a scene is a key issue for autonomous mobile robots in crowded areas. Therefore, accurately tracking people from a mobile platform can help improve interaction effectively and efficiently. Unfortunately, tracking in a real environment is extremely difficult because the non-rigid human's poses, variant appearances, and cluttered occlusion. So in recent years, a lot of new strategies have been proposed for robot tracking. Rui et al. [7] achieved a robotic tracking system based on mobile robot vision using adaptive color matching and Kalman filter. Kim et al. [8] proposed that present object as a point, and utilized particle filter to detect the position of object, then the range of the object is calculated by laser range finder and the finally position is obtained. Lang et al. [9] proposed a method that used the SIFT to achieve robust object tracking, then the image-based visual servo (IBVS) controller drives the robot toward the object in the feedback loop. From those papers, we can know that the visual tracking algorithms are key issue for object tracking system.

For most exiting tracking algorithms, a robust appearance model is crucial for avoiding the failure of whole tracking model. So according to different appearance representation schemes, the object tracking algorithm can be categorized as two aspects: generative models and discriminative ones. Generative models are directly designed to find candidates which are most similar to the object templates in each frame. Ross et al. [10] introduced a tracking approach that used incremental subspace to efficiently represent the online change of object appearance. Based on the sparse representation, Mei et al. [11] proposed a method to deal with the partial occlusion by linearly combining arbitrary templates. Zhang et al. [12] integrated compressive sensing theory in a real-time tracking framework and demonstrated that object samples represented by low dimensional samples could be excellent performance. However, these generative models do not take into account background information. It is difficult to keep excellent performance when similar objects exist nearby the object.

Otherwise, discriminative appearance models distinguish the object from the background by training a robust classifier, namely, find the optimal decision boundary to division whether it is object samples or not. So these tracking-by-detection methods are mainly depend on the quality of the train dates. Avidan et al. [13] proposed an off-line tracking method based on support vector machines and used it in conjunction with the optical flow classifier to detect vehicles from the background. Based on the boosting method, Avidan et al. [14] used the weighting of each sample to train the weak classifier and integrated the strong classifiers. Collins et al. [15] obtained the features by the on-line form and showed that this method improved the confidence of the classifiers, but this method had the problem of self-learning which will accumulate when there are errors in the current

tracking results, consequently, the model often drifts, eventually leading to tracking failure. In order to reduce the error accumulation, Grabner et al. [16] proposed a semi-supervised online boosting discriminant tracking method, combined with the first frame samples obtained from the prior classifier and on-line classifiers. To further address the drift problem, Babenko et al. [17] proposed an on-line multiple instance learning tracking method that could avoid the ambiguity of the definition of positive and negative samples by putting multiple instances into one bag and labeling the bags. Zhang et al. [18] enhanced the tracking effect by assigning different weights to the samples in the bag. However, in the multiple instance learning, the greedy algorithm is used to maximize the feature of a log likelihood function. The selected feature does not necessarily contain high information content, so a large number of samples have to be selected to achieve discriminant effect. Therefore, in the discriminant model tracking algorithms, some high-quality sample sets can effectively improve the performance of the classifier. The general discriminant models collect samples are done by sampling and labeling. The sampling process generates a set of samples around the current tracking result, and labeling these samples relying on heuristics strategy. But this example collection method employs a separate component for managing the training set and ignores the correlation between sampling and labeling process, so the selected samples by this strategy are not most informative or useful for classifier training, which can't improve the performance of classifiers and may cause drift [19, 20, 21].

The tracking algorithm based on deep learning [22, 23] uses multi-layer network structure to extract essential features of the images, which can effectively achieve expression of image. A large number of natural image sets are used to pre-train deep network parameters through unsupervised features learning, and the optimized network structure is applied to online tracking by transfer learning, which can effectively solve the problem that erroneous tracking and target drift caused by insufficient training samples under complex conditions. However, since the target information constantly changes in different time during the tracking process, how to obtain the optimal feature set in the deep network is a problem that needs to be solved at present. When the number of features is too large, the target model is over fitting or converges too slowly. In contrast, when the number of features is too small, the model would be caused under fitting with lacking relevant feature. In tracking algorithm, only the sigmoid function is used as the classification model. The classification model is simple, but the problem of misclassification often occurs in complex background.

The online multiple instance learning (MIL) target tracking algorithm uses plurality samples forming positive package selected from the tracking result of previous frame to update appearance model, which avoids an erroneous update due to error accumulation of the previous frame, and when the strong classifier judges the new test sample whether or not is the target location, it is obtained by calculating the

probability likelihood function of the package. All are also favorable for comparing the similarity of the matching and the candidate target sample from the perspective of the entire image sample, and can also increase success rate of tracking to a certain extent. However, simultaneously, the multiple instance learning tracking method assumes that the distribution of samples in the positive packet is the same, that is, the importance of the samples is not distinguished. This will, to a certain extent, easily lead to the accumulation of errors when the classifier is updated, and gradually expands and eventually leads to tracking drift. In addition, when using a greedy algorithm to maximize a log-likelihood function to obtain samples in multiple instance learning tracking algorithm, the selected sample does not necessarily have high information content.

In order to design a human detection and tracking system, we present an improved multiple instance learning based on active learning with co-training framework. The main contributions of this study are as follows.

(1) Because the multiple instance learning algorithm has positive and negative samples in the positive bag, the sample selection in the MIL is through the optimization of the bag likelihood function, so that the selected samples have no purpose and do not consider the classifier training. In this paper, by optimizing the Fisher information matrix, we can make the selected samples more informative, and improve the discriminative ability of the classifier.

(2) Because of the self-training problem in multiple instance learning, we use two classifiers which are trained in two views to complete the cooperative training, and ensure the accurate updating of the joint classifiers.

(3) In this paper, the proposed improved multiple instance learning based on active learning with co-training framework is conducted test classic vide, which verified the robustness and real-time performance of the proposed algorithm. Then the target tracking method proposed is validated on MT-R robot of the wheeled mobile. The experimental results show that this method has strong robustness for human target tracking when they move and their appearance change.

This paper is organized as follows. In the next Section, we review the related work. The Section 3 provides implementation details of the proposed tracker, and the framework of our robot tracking system is shown in Section 4. The detailed experiment setup and results are represented in Section 5. The conclusion is given in Section 6.

2. Related works

2.1 Online multiple instance learning tracking method

Multiple instance learning has been widely used in visual tracking [17, 18]. Suppose that we

have a training data $\{(X_i, y_i)\}_{i=1,2,\ldots,n}$, where $X_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ and $y_i \in (0,1)$ is the label of each bag. Therefor a bag is positive if it contains at least one positive instance; otherwise it is negative. In the multiple instance learning tracking algorithm, the object position of the first frame is known, and the optimal classifier is selected by using the gradient descent algorithm to maximize the likelihood function,

$$L = \sum_i \left(y_i \log p(y_i = 1 | X_i) + (1 - y_i) \log \left(1 - p(y_i = 1 | X_i)\right)\right) \tag{1}$$

Which is based on the package on the weak classifier selection strategy, rather than in the sample selection, because in the training bag within the sample label is not known, according to the Noisy-OR model, estimate the probability characteristics of bag level,

$$p\left(y_i = 1 | X_i\right) = 1 - \prod_j \left(1 - p\left(y_i = 1 | x_{ij}\right)\right), \quad p\left(y_i = 0 | X_i\right) = \prod_j \left(1 - p\left(y_i = 1 | x_{ij}\right)\right) \tag{2}$$

Where $p\left(y_i = 1 | x_{ij}\right)$ is predicted by strong classifier $H_K\left(\boldsymbol{x}_{ij}\right)$:

$$p\left(y = 1 | \boldsymbol{x}_{ij}\right) = \frac{\exp\left(H_K\left(\boldsymbol{x}_{ij}\right)\right)}{\exp\left(H_K\left(\boldsymbol{x}_{ij}\right)\right) + \exp\left(-H_K\left(\boldsymbol{x}_{ij}\right)\right)} \tag{3}$$

We use feature vector $f_k\left(\boldsymbol{x}_i\right) = \left(f_1(x), \ldots, f_k(x)\right)$ to represent Haar-like feature of example x, suppose $f_k(x)$ are independent, $p(y=0) = p(y=1)$ and $p(f_k(\boldsymbol{x}) | y_i)$ obey Gauss distribution, we can get under Bayes rule:

$$h\left(\boldsymbol{x}_i\right) = \log\left(\frac{p(f_k(\boldsymbol{x}) | y = 1)}{p(f_k(\boldsymbol{x}) | y = 0)}\right) = \sum_{k=1}^{K} \lambda_k h_k\left(\boldsymbol{x}_i\right) \tag{4}$$

The MIL tracker first maintains a pool of M weak classifiers as $\Phi = \{h_1, \ldots, h_M\}$, and greedily selects K< M weak classifiers from Φ via the following criterion:

$$h_k = \arg\max_{h \in \Phi} L\left(H_{k-1} + h\right) \tag{5}$$

Where $H_{k-1} = \sum_{j=1}^{k-1} h_k$ is a linear combination of the first k-1 weak classifiers, and L(.)

is denoted by formula (1).

2.2 Active learning

Active learning theory [24] is designed to select the unlabeled sample. It contains the optimal amount of information. The core idea is to take into account that different samples have different effect on the final classifier actually in the machine learning, namely, the amount of statistical information. The bigger the sample size is, the more important the classification interface will be confirmed. The so-called sample information is relative to the uncertainty of the classifier, with the sample set S, trained on the training data to obtain the concept category set C, in other words, each of the concept categories in C can correctly determine the training data.

In the classifier learning and sample selection, there will be inconsistencies, resulting in the selected samples don't produce enough effect on the update of the classifier, active learning can improve the consistency of the two parties. It also makes the selected sample more representative and the corresponding reduction in the number. Active learning will count actively and select the characteristics which contain the useful information.

3. Proposed algorithm

3.1 Online example extraction of proposed tracker

In this paper, we use the features which are called as color histograms and HOG to describe the frame of the video sequence respectively. We obtain two fully redundant views and cooperate with the two views to training classifiers. Color histograms focus on deformation because they are insensitive to the position change of image pixel, while HOG features pay more attention to edge change information (*i.e.*, robustness to illumination changes and so on). Therefor we combine those two kinds of features to distinguish the easily confused samples and correct a better ability.

3.2 Online multiple instance learning object tracking algorithm based on active learning

In the multiple instance learning tracking algorithm, the positive sample bag contains both positive and negative samples. In the formula (1), the selection classifier is obtained by maximizing the logarithmic function of the positive sample set. The strategy does not take into account the following classifier training problems, it do not choose the most useful samples to train the classifier, which will lead to the instability of the classifier. In order to the more accurate selection and the confirmation of the object location, based on the active learning, we propose an active example selection method. In this paper, we introduce the Fisher matrix into the formula (1) to minimize the variance of the classifier, so that the selection of the good samples and reduction of the volumes in the features required for

training. Fisher matrix is commonly used to measure sample probability in active learning. Minimizing the predictive variance of the classifier with the substituted sample, it is also effective to eliminate the confusing samples.

In the formula (3), the strong classifiers are composed by the weak classifiers:

$$H(\mathbf{x}) = \sum_{k=1}^{K} \lambda_k \tilde{h}_k(\mathbf{x}) = \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{h}(\mathbf{x}) \tag{6}$$

Where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)^{\mathrm{T}}$ refers to each of the weak classifiers, and $\boldsymbol{h} = (h_1, \ldots, h_k)$ denotes the corresponding weight of weak classifiers. In our case, we aim to select the sample subsets which contained the most informative to train the classifier, and to find the corresponding weight $\boldsymbol{\lambda}$. According to the theory [25], due to conditional independence of the $p(y|\boldsymbol{\lambda})$, the following formula about Fisher information matrix $\mathrm{I}(\boldsymbol{\lambda})$ can be obtained:

$$\mathrm{I}(\boldsymbol{\lambda}) = -\int p(X_i) p(y_i|X_i) \frac{\partial^2}{\partial \boldsymbol{\lambda}^2} \log p(y_i|X_i) dX \tag{7}$$

Where $p(X_i)$, it represents the probability of the bags that are labeled already. The Fisher information matrix quantifies the amount of information that an observation carries about an unknown parameter. So we measure the information of the samples by establishing a Fisher information matrix on the samples based on the bags:

$$\mathrm{I}(\boldsymbol{\lambda}) = \sum_i \left[y_i p(y_i|X_i) \frac{\partial^2}{\partial \boldsymbol{\lambda}^2} \log p(y_i|X_i) + (1-y_i) p(y_i|X_i) \frac{\partial^2}{\partial \boldsymbol{\lambda}^2} \log p(y_i|X_i) \right] \tag{8}$$

Where $p(y_i = 1|X_i) = 1 - \prod_j \left(1 - \sigma\left(\boldsymbol{\lambda}^T h\left(\mathbf{x}_{ij}\right)\right)\right)$, $p(y_i = 0|X_i) = \prod_i \left(1 - \sigma\left(\boldsymbol{\lambda}^T h\left(\mathbf{x}_{ij}\right)\right)\right)$.And to obtain the most informative samples, we have to get suitable $p(X_i)$ maximize the trace of $\mathrm{I}(\boldsymbol{\lambda})$ [26], so $tr\left(\mathrm{I}(\boldsymbol{\lambda})\right)$ is represented by:

$$tr\left(\mathrm{I}(\boldsymbol{\lambda})\right) = -\frac{1}{2} \sum_i \left[tr\left(\sum_{y_i \in \{0,1\}} y_i p\left(y_i|X_i\right) \frac{\partial^2}{\partial \boldsymbol{\lambda}^2} \log p\left(y_i|X_i\right) \right) \right] \tag{9}$$

Because of each weak classifier $\boldsymbol{h} = (h_1, \ldots, h_k)$ is stump function, so we have

$\boldsymbol{h}\left(x_{ij}\right)^T \boldsymbol{h}\left(x_{ij}\right) = k$. They are introduced into the above formulas:

$$tr\left(\mathrm{I}\left(\boldsymbol{\lambda}\right)\right) = \frac{k}{2}\sum_i \left(\sum_j p\left(y_i = 1 \middle| x_{ij}\right)^2 \frac{\prod_l \left(1-\sigma\left(\boldsymbol{\lambda}^T h\left(\boldsymbol{x}_{ij}\right)\right)\right)}{1-\prod_l \left(1-\sigma\left(\boldsymbol{\lambda}^T h\left(\boldsymbol{x}_{ij}\right)\right)\right)} \right)$$

$$= \frac{k}{2}\sum_i \left(\sum_j \left(\frac{1}{1-\prod_j \left(1-\sigma\left(H\left(x_{ij}\right)\right)\right)} - 1 \right) \sigma^2\left(H\left(x_{ij}\right)\right) \right) \tag{10}$$

Finally, in our method, though maximize the probability $p\left(y_i = 1 \middle| x_{ij}\right)$ and $p(y_i = 1 | X_i)$ of positive and negative bags respectively, the fisher information matrix could select samples which contain the most of the information.

The overall flow chart of this paper is shown in **Fig. 1**. Assume that the target position $l_t\left(\boldsymbol{p}\right)$ in t time is known, what the samples are extracted from $X^+ = \left\{ \boldsymbol{p} \middle| \left\| l_t(\boldsymbol{p}) - l_t(\boldsymbol{p}^*) \right\| \le a \right\}$ range would be constituted positive pack, what are extracted from $X^- = \left\{ \boldsymbol{p} \middle| a \le \left\| l_t(\boldsymbol{p}) - l_t(\boldsymbol{p}^*) \right\| \le b \right\}$ would be constituted negative pack, where $a < b$, and $\boldsymbol{p}^*$ is target sample. Afterwards, $\left(X^+, X^-\right)$ is used to update classifier. When $t+1$ frame is coming，the sample is collected in $X^s = \left\{ \boldsymbol{p} \middle| \left\| l_{t+1}(\boldsymbol{p}) - l_t(\boldsymbol{p}^*) \right\| \le s \right\}$, We choose K weak classifiers based on the proposed active sample selection strategy from M weak classifier pools. The corresponding $h_k\left(p\right)$ and $H_k\left(\boldsymbol{p}\right)$ are calculated with these weak classifiers, and combine the collaborative training to obtain the target location. The complete procedure of the improved MIL tracking method is shown in **Algorithm 1**.

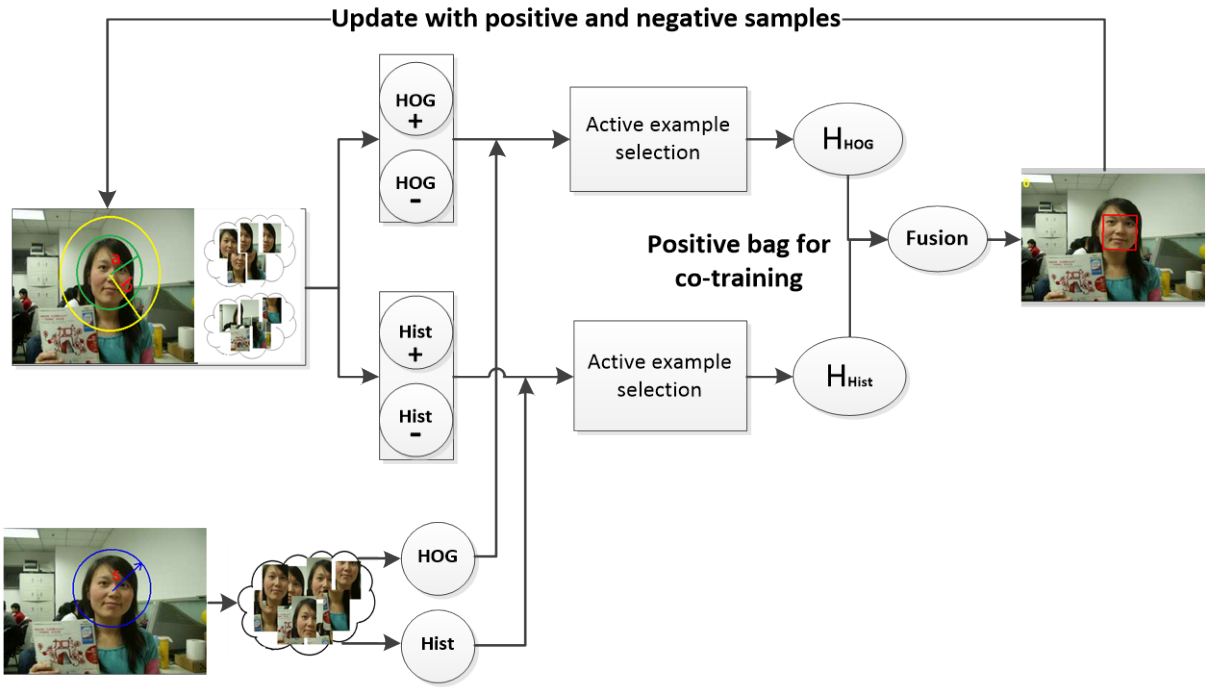


Fig. 1. The overview of the improved MIL algorithm

Algorithm 1 The improved MIL tracking algorithm

Input: Training set: $\{X_i, y_i\}_{i=1}^{N+L}$, $X_i = \{x_{i1}, x_{i2}, \cdots,\}$, $y_i \in \{0,1\}$

1. Update all weak classifier in the pool
2. Initialize $H_0(x_{ij}) = 0$
3. **for** $k = 1$ to K **do**
4. **for** $m = 1$ to M **do**
5. $L_m = L(H_{k-1} + h_m)$
6. **end**
7. $m^* = \arg\min L_m$
8. $h_k \leftarrow h_m^*$
9. $H_k = H_{k-1} + h_k$
10. **end**

3.3 Classifier update of co-training framework

In order to predict the possible position of the object, the arrival of the new video frame, acquisition of positive and negative samples, with a frame of the trained HOG classifier and color histograms respectively for these samples are classified to obtain the classifier corresponding to sample values of confidence, and confidence is finally linear weighted. To update the classifier, the co-training strategy use a small amount of training samples labeled two initial classifier, and then in the learning process, these classifiers select several high confidence level of the unlabeled samples and used to update the other classifier, this method effectively improve the classification performance. Two prerequisites for co-training: 1) If the training sample is sufficient, a strong classifier can be learned on each feature set; 2) Each feature set is independent of another feature set when given a class marker, co-training algorithm can effectively use unlabeled samples to enhance the performance of the classifier [27, 28, 29, 30]. The co-training formula:

$$p^{HOG}\left(y=1\middle|x_{ij}\right)=\frac{1}{1+e^{-H_k^{HOG}\left(x_{ij}\right)}},\ p^{Hist}\left(y=1\middle|x_{ij}\right)=\frac{1}{1+e^{-H_k^{Hist}\left(x_{ij}\right)}} \tag{11}$$

$$\omega^{Hist}=\frac{p^{Hist}\left(y=1\middle|x_{ij}\right)}{p^{Hist}\left(y=1\middle|x_{ij}\right)+p^{HOG}\left(y=1\middle|x_{ij}\right)},\ \omega^{HOG}=\frac{p^{HOG}\left(y=1\middle|x_{ij}\right)}{p^{HOG}\left(y=1\middle|x_{ij}\right)+p^{Hist}\left(y=1\middle|x_{ij}\right)} \tag{12}$$

$$co\_x_{ij}=\arg\max_{x\in\left\{X^s\right\}}\left(\omega^{Hist}p^{Hist}\left(y=1\middle|x_{ij}\right)+\omega^{HOG}p^{HOG}\left(y=1\middle|x_{ij}\right)\right) \tag{13}$$

Where ω^{Hist} and ω^{HOG} is the weight of color histogram and HOG respectively. Moreover, when the HOG classifier do well in the regions that the color histogram classifier performance poorly, the output will rely heavily on ω^{HOG}, conversely, the ω^{Hist} will occupy a large part of outcome. Outputs $co\_conf_k(x)$ with higher confidence value can be used as positive samples. And it is apparent that instances with high confidence value contain the most informative instance. Then we use these positive instances form a positive bag, and to training the other classifiers.

In order to characterize the appearance change of target in real time, the classifier needs to be updated. Firstly, after the target position is determined, samples are collected at a fixed distance around it to form positive packet and two classifiers are updated. For negative

sample, samples with a confidence level in the middle value are selected, samples are collected to form negative packets, and they trained each other, which achieve that they can give samples that are more difficult to handle to each other for handling. If these samples appear in next time, it can rely on the results of other classifier to increase robustness in system. An overview of the proposed tracking methods is summarized in **Fig. 2** and **Algorithm 2**.

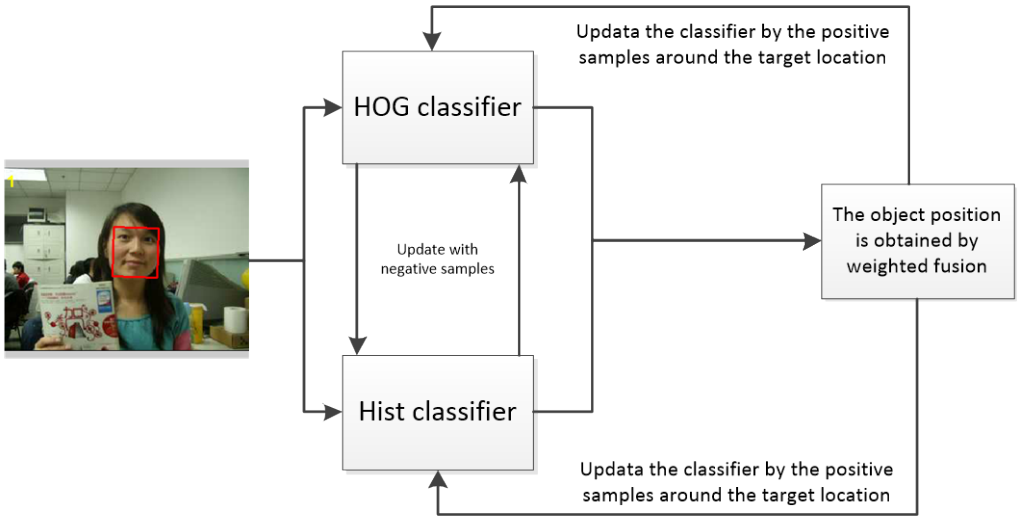


Fig. 2. The framework of co-training strategy

Algorithm 2 The proposed tracking algorithm based on co-training framework

Input: t-th video frame

Output: Object location: $H(x)=\sum_{k=1}^{k}\lambda_k h_k(x)$

1. Initialization: the location of object in first frame is known, and in the radius of the neighborhood to extract a number of image blocks as a positive bag, $X^{+}=\left\{x\left\|l_t(x)-l_{t-1}^{*}(x)\right\|\le a\right\}$ and $X^{-}=\left\{x\left|a\le\left\|l_t(x)-l_{t-1}^{*}(x)\right\|\le b\right.\right\}$; The samples are represent based on color histograms and HOG respectively, then use these feature to calculate classifier parameters and constructed a weak classifier pool;

2. From second frame, crop out a set of image patches $X^{+}=\left\{x\left\|l_t(x)-l_{t-1}^{*}(x)\right\|\le R\right\}$, where R<α , $l_{t-1}^{*}(x)$ is the last location of object, and the color histograms and HOG are also extracted

respectively;

3. According to the active example selection method proposed in Algorithm 1, we find most discriminative weak classifiers. Then based on combined outputs, the final position $l_t^*(x)$ is obtained by formula (11), (12) and (13).

4. Acquire new samples forming new bags from each classifier by using co-training framework and extract their features to update the other classifiers.

4. Object tracking system of moving robot

In this paper, based on the proposed object tracking algorithm and mobile robot MT-R, an object tracking system of moving robot is presented, and this system is divided into three parts: system initialization, visual object tracking and robot motion control. The following is a brief introduction to the main functions of these three parts:

(1) System initialization: The initialization of the mobile robot, including the interface parameters and the image acquisition of the camera, and the communication among the modules.

(2) Visual object tracking: This part is mainly realized by the improved algorithm running on the computer. First through the OpenCV function library using C++ program in the object detection template to search in the global, while the object is detected, this frame is used as initial frame. Then in two views, color histogram features and gradient histogram features are built for the classifier training, combined with the co-training strategy to train a robust classifier, used in the next frame to determine the object location, and real-time calibration and display in the image.

(3) Robot motion control: After getting the position of object in the image, the moving strategy of robot is obtained, including the steering angle, acceleration and distance. Through the serial bus to the robot motion control module to ensure that the object near the center of the location. Finally, the position of object tracking is obtained.

5. Experimental results and analysis

In this part, the proposed tracker are tested on 10 video sequences of the OTB-13 [31] to verify the advantages of our algorithm, these sequences contain the occlusion, rotation, illumination changes, fast movement, deformation and other complex situations. At the same time, we compared the other six state-of-the-art algorithms on the same sequence: IVT[10], CT[12], MIL[15], Co-mil[29], TLD[32], Struck[33]. We set the size of the initial tracking rectangle to 32×32 pixels. In addition, the HOG features are extracted by 18×18 patches on 9 blocks with step length of 7 pixels, so the features are represented by a 288-dimensional feature vector. In this paper, we select $K=20$, our tracker selects 20

features for classifier construction and has highly efficient. The number of positive samples is 30, the number of negative samples is 60. α = 20, β = 25, negative samples were collected in a ring region with radius of 20 to 25, and positive samples in each frame were sampled in range of radius γ = 4. We find that proposed algorithm is quite robust. The search range S = 30, which is enough to consider all possible target positions, because the target motion between two consecutive frames is usually smooth. The learning rate η = 0. 85, a smaller learning speed allows the tracker to quickly adapt to rapid changes in appearance, and a larger learning rate can reduce the possibility of tracker drifting away from the target. It can be obtained robust results by fixing η = 0. 85 in our experiments.

5.1 Quantitative analysis

The experiment uses the center error and the average overlap rate to measure the merits of the tracking algorithm. The center location error can reflect the overall stability of algorithm, it is Euclidean distance between center location of each frame obtained by this algorithm and the center location of actual picture. The calculation formula is as follows:

$$CLE = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

(14)

Where, x_i and x_j represent the coordinate values of tracking results on x axis and y axis, x_j and y_j represent the actual coordinate values of target on x axis and y axis. From the formula, we can see that center location error is the smaller, the better. This paper considers the average error of the center location to be 20 pixels is tracking success.

While the overlap rate can reflect accuracy of tracking. The definition of the overlap rate is:

$$score = \frac{area(ROI_t \cap ROI_g)}{area(ROI_t \cup ROI_g)} \tag{15}$$

Where, ROI_t is algorithm tracking frame, ROI_g is really value frame. The final calculated result $score$ can measure the accuracy of algorithm. Overlap ratio is the greater, the better, and ideally it is 1. Under the criterion of the average overlap ratio, if the average overlap ratio of $score$ in a certain sequence is greater than 0.5, the tracking can be considered as successful. It can be seen from **Table 1** and **Table 2** that this algorithm has achieved some results in these ten sequences.

Table 1. Center Location Error. The best two results are shown in red, blue.

Sequence	MIL	CT	IVT	TLD	Co-mil	Struck	Ours
Deer	72.14	88.55	107.62	31.21	18.61	12.15	10.05
CarDark	55.60	19.12	1.96	16.78	4.60	1.48	1.67
David	17.90	12.61	4.82	8.40	22.68	52.82	4.90
Lemming	56.20	35.44	120.21	84.60	22.12	19.80	24.89
Mhyang	15.12	8.80	3.80	5.10	12.24	2.90	3.82
Walking	8.70	9.60	5.30	13.32	3.21	4.61	4.42
Fish	24.21	22.30	7.60	12.53	5.87	4.38	4.05
Girl	42.06	24.40	36.28	28.32	19.37	12.30	15.37
FaceOcc2	22.32	19.06	14.40	18.56	7.16	6.67	6.50
FaceOcc1	32.21	38.14	23.02	33.80	15.31	21.60	13.80

Table 2. Overlap Rate. The best two results are shown in red, blue.

Sequence	MIL	CT	IVT	TLD	Co-mil	Struck	Ours
Deer	0.38	0.07	0.25	0.49	0.27	0.70	0.72
CarDark	0.28	0.18	0.48	0.54	0.10	0.41	0.67
David	0.44	0.50	0.64	0.56	0.65	0.33	0.58
Lemming	0.45	0.36	0.25	0.55	0.28	0.57	0.74
Mhyang	0.66	0.69	0.52	0.76	0.72	0.80	0.74
Walking	0.45	0.52	0.67	0.45	0.54	0.58	0.66
Fish	0.45	0.46	0.74	0.76	0.53	0.82	0.83
Girl	0.42	0.60	0.26	0.70	0.51	0.79	0.76
FaceOcc2	0.59	0.59	0.68	0.52	0.71	0.65	0.77
FaceOcc1	0.60	0.63	0.66	0.49	0.63	0.75	0.64

5.2 Qualitative analysis

As shown in **Fig. 3**, there are results of six tracker and our proposed tracker on ten sequences (MIL: green, CT: grey, TLD: blue, IVT: sky blue, Co-mil: wine, Struck: pink, our proposed tracker: red). This part we mainly discuss the four challenges below:

1) *Occlusion and motion blur*: As shown in **Fig. 3(i)** and **Fig. 3(j)**, *Faceocc1* and *Faceocc2* are all human face that occluded partially. In the *Faceocc2* sequence, the object also undergoes head rotation that confuse the online update in the TLD and CT methods and cause the drift problem. Although the MIL and CT tracked the object successfully, the accuracies are far from satisfaction. Only the Struck and our method are able to keep the track on the object and obtain favorable results. In the *Deer* sequence, the deer moves faster at 5#. CT, TLD and MIL are all draft away. Then the object suffers from the cluttered background, the several tracker loss the object eventually because the error calculation. The reason for the satisfied performance of our method is that it takes full advantages of

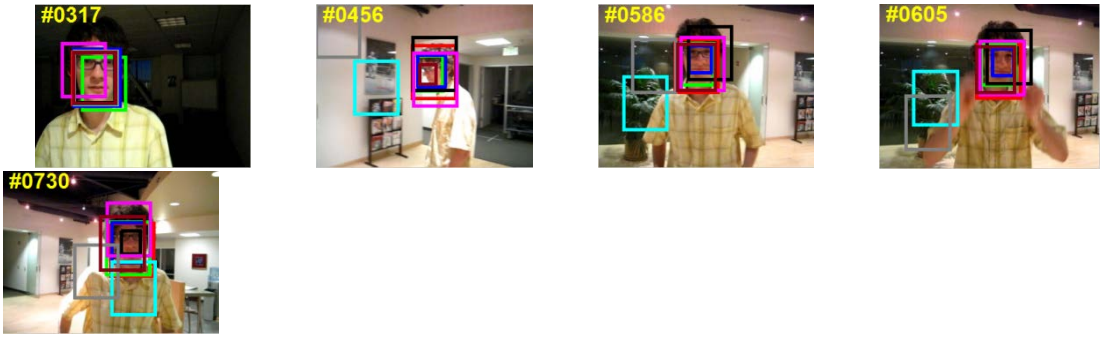
co-training which collaborating two individual classifiers to training that utilizes more information.

2) *Illumination Variations*: In *David* sequence, the object suffer from illumination and background change when the man walk and turn around. As shown in **Fig. 3(a)**, when the object appearance change drastically at 456#, the CT and Struck loss the object. Only IVT and our methods obtain favorable tracking result. **Fig. 3(c)** shows the experimental results of seven algorithms on *fish* sequence, the object suffers from illumination and mobile camera. MIL and CT gradually loss the object when the camera moves to the right direction. The Struck and our method achieve satisfied performance. In this sequence, we always keep track of the object because that color histogram is insensitive to the change of illumination.

3) *Clustered background*: In *lemming* sequence, the background of object is clustered, and the object is also almost occluded for a while, so the performance of the classifier is obviously decreased, which leads to the loss of CT and COMIL at the beginning of the sequence. In the *CarDark* sequence, the contrast between car and background is very low, the MIL, TLD and CT algorithms loss the object in the beginning, so the error is far greater than the tracker proposed in this paper, which shows that the strong classifier composed of the weak classifier based on the active learning model has stronger ability of discrimination.

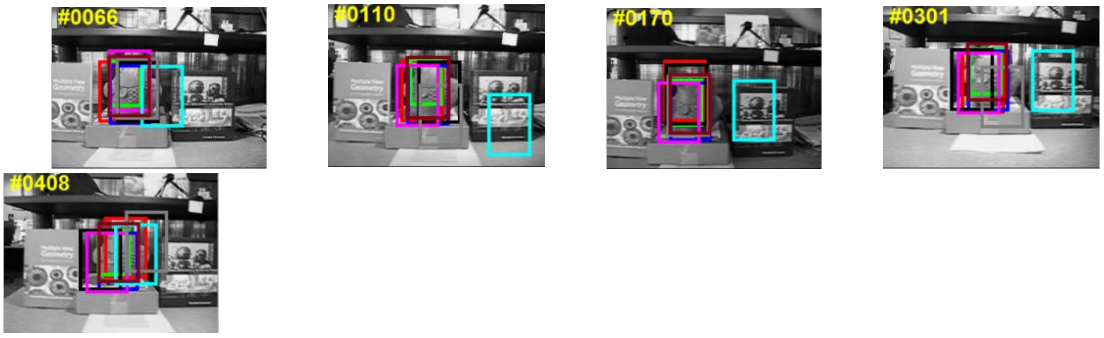
4) *Deformation*: In the *Myhang* sequence, the person changes his expression and viewpoint frequently, moreover, he also suffers from illumination changes on his face. The MIL, IVT and CT methods don’t achieve well. In the *Walking* sequence, the walker look small, and as he walks, his body is getting change. Several tracker lose the object when the walker occluded by streetlight and another person. The IVT and our method achieve more stable performance in the entire sequence. When the object suffers from deformation, our method can select more discriminative features via active example selection.

The packet probability likelihood loss function is replaced by optimizing the Fisher information matrix in this paper, which can make the selected sample has more informative and improve the discriminant ability of the classifier. Then, cooperative training strategy is introduced, and two multiple instance learning classifiers are established by training histogram of gradient features and color histogram features. The final target position is obtained by merging the tracking results of two classifiers, and the multiple positive samples around tracking result is extracted during the update process, which form positive packet. The sample with the confidence in middle position is extracted as negative packet. Then the classifier is updated with the positive and negative sample packets. This strategy can reduce the impact of previous frame's tracking error to minimum, which guarantees accurate update of the joint classifier. This ensures that the algorithm has high tracking accuracy in occlusion, motion blur, clustered background and deformation.

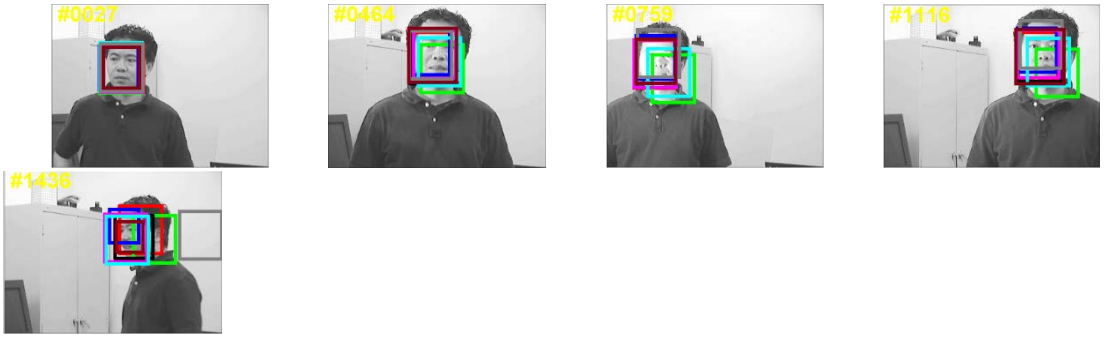


(a) David

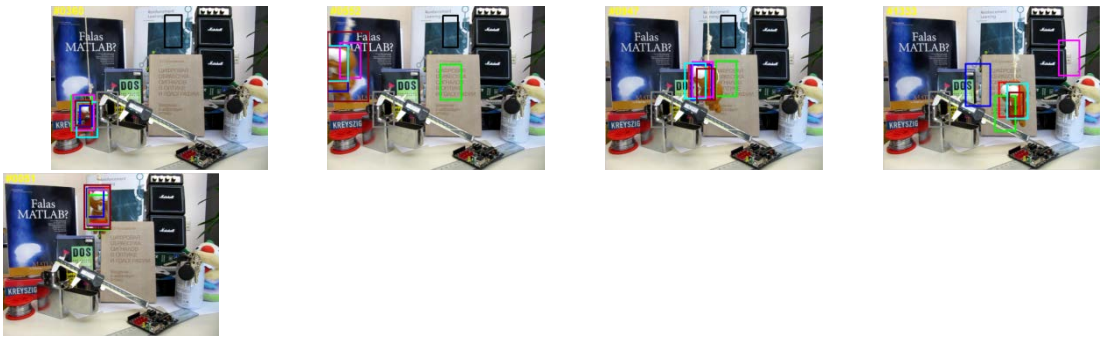
(b) CarDark



(c) Fish

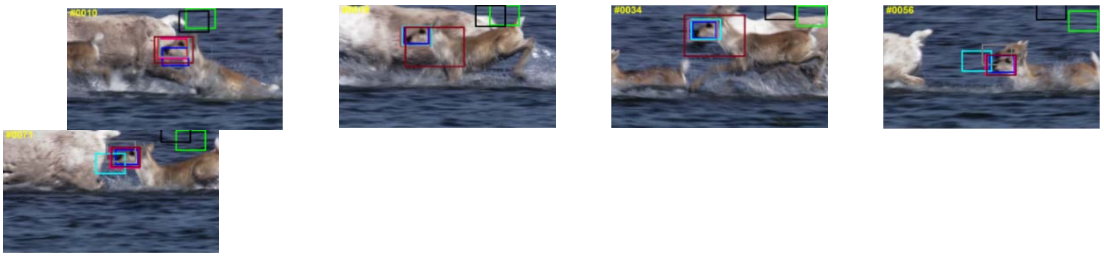


(d) Myhang

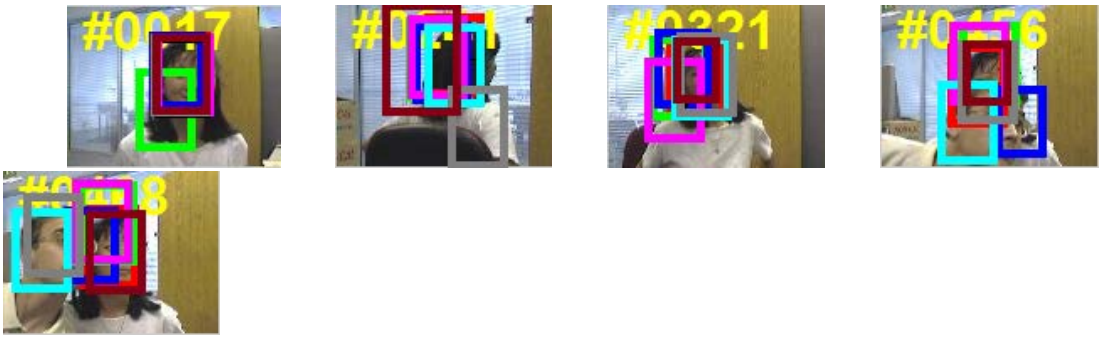


(e) Lemming

(f) Walking

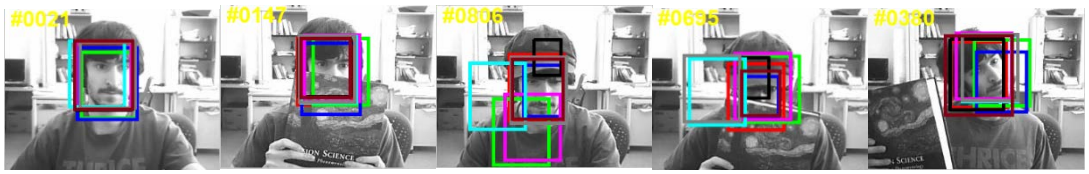


(g) Deer



(h) Girl

(i) Faceocc1



(j) Faceocc2

Fig. 3. Screenshots of tracking results

5.3 Robot tracking experiment

In order to validate the robustness of proposed algorithm under different situation, the moving robot is experimented under cluster background, occlusion, illumination changing etc. In this experiment, the object is defined as human face. After the initialization of the robot system, the robot vision target tracking search globally through the object detection template in OpenCV function library, and a robust classifier is trained in conjunction with cooperative training strategy to determine target position in next frame. After obtaining the pixel position of target in image frame, the robot's movement strategy is got by calculation. Experiments are carried out in different scenarios, each experiment with two rows of images which represent the tracking results and the moving robot respectively.

The following experiments are conducted in different scenarios：

1) *Occlusion*： As shown in **Fig. 4**, the tracking results of the robot in the case of serious occlusion of the object are taken from 279 frames, 280 frames, 281 frames and 288 frames respectively. As can be seen in the graph, the full occlusion of the face occurs on the three consecutive frames. This makes it possible to mistake contaminated samples as labeled samples, which can lead to unstable performance of the classifier. And according to the results of experiment, the robot keeps tracking the object while the object is gradually occluded. This is mainly due to the fact that the strategy of the active sample selection increases the correlation between the sample selection and the classifier training, so that the classifiers also ensure that the robot can always track the object in the case of less effective samples.

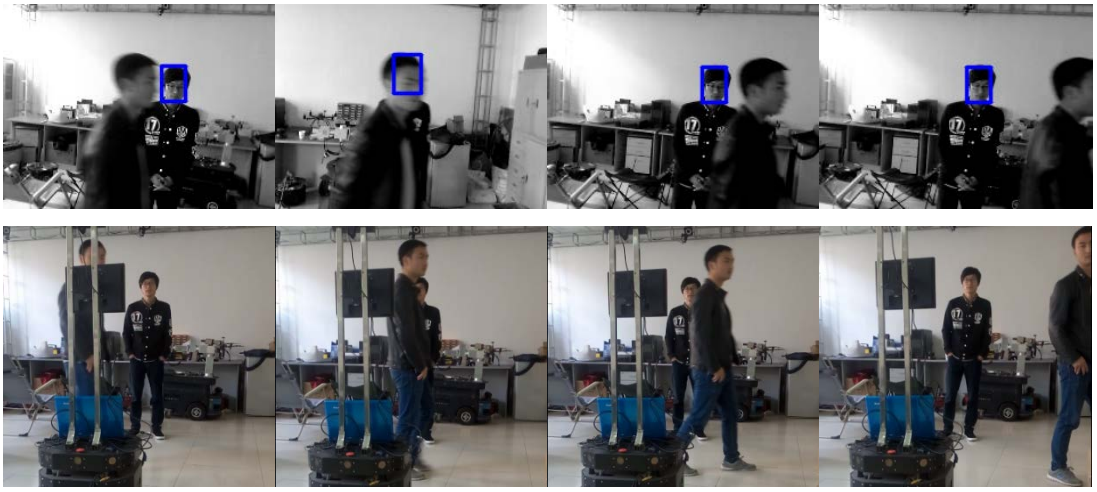


Fig. 4. Robot tracking under occlusion

2) *Illumination change*: The **Fig. 5** shows the tracking result of robot in the case of light changes, the challenge of this group is that when the object move around, the light changes obviously, so it is more and more difficult to distinguish the object and background. From the tracking results, we can see the robot has maintained an accurate track of the object, and the robot is also following the object by the rotation, because the HOG feature in this paper has good robustness to illumination change, when the Hist feature does not work under the illumination change, the strategy of co-training allows HOG features to be given more weight, so the final result is still good.

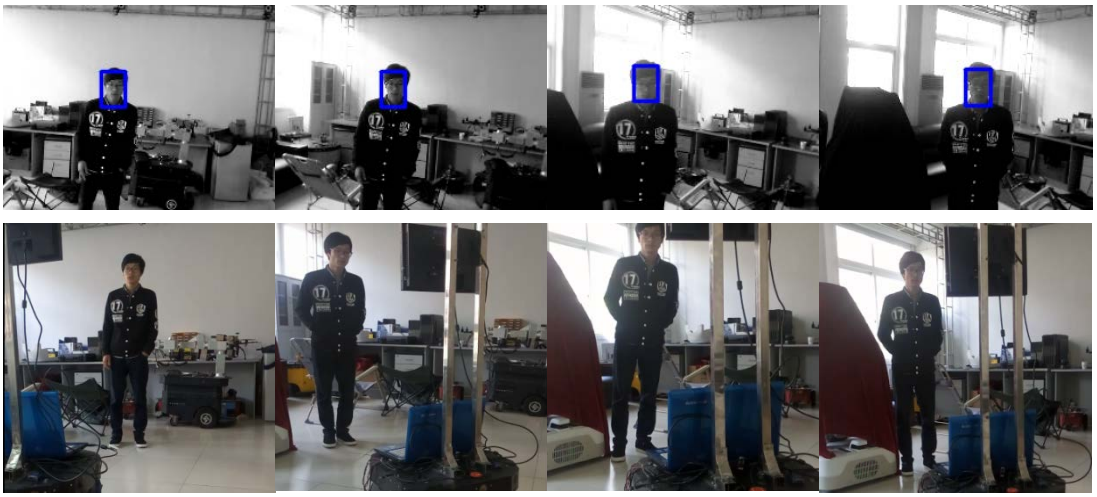


Fig. 5. Robot tracking under Illumination change

3) *Rotation and scale changes*: As shown in **Fig. 6**, the robot tracking the object under the situation of rotation and scale changes, it is 190 frame, 204 frame, 296 frame and 323

frame. In the case of object rotation, the object appearance changes greatly, but our method is based on the discriminative methods which can effectively distinguish the object and background. While the object move vertically, the robot moving at the same time, so the proposed method solves the problem of object rotation and controls the motion of the robot following the object.

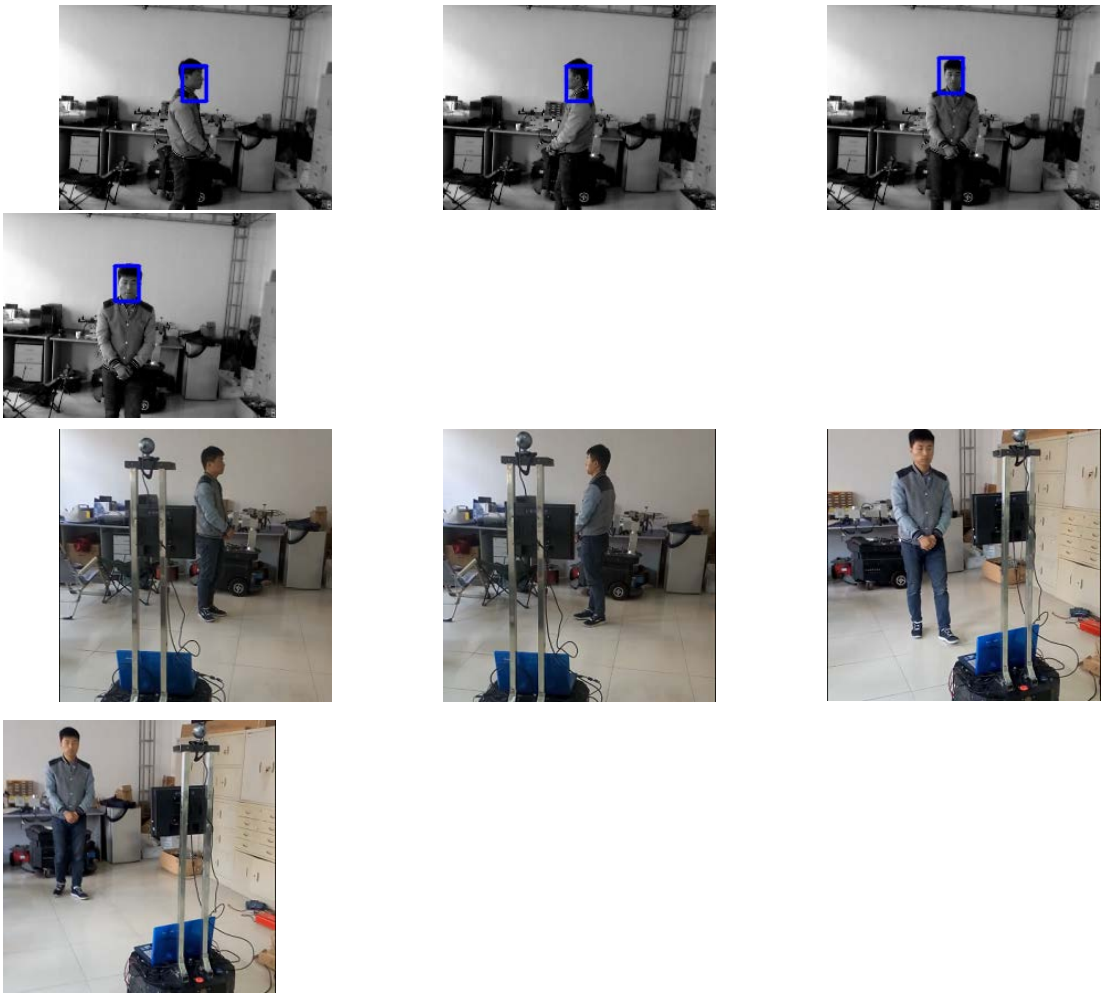


Fig. 6. Robot tracking under fast moving object

4) *The comparison of our method and MIL tracker under occlusion*: As shown in **Fig. 7** and **Fig. 8**, the tracking experiments of the proposed algorithm and the MIL tracking algorithm in case of occlusion of the object are respectively presented. And this two sequences have occlusion in a series of frame, and it's apparently that our method perform well than MIL tracker. It is because that the active example selection strategy in our tracker is more effect than the MIL tracker, the traditional online MIL tracking algorithm can add the error samples collected into occlusion to the labeled samples, and affect the training of

classifier, so it is easy to drift when the object is completely occluded.

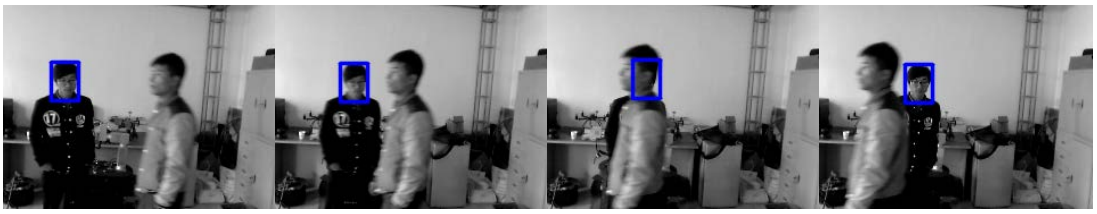


Fig. 7. The proposed method

Fig. 8. MIL tracker

5) *The comparison of our method and MIL tracker under pose variation*: As shown in **Fig. 9** and **Fig. 10** the tracking experiments of the algorithm and the MIL tracking algorithm in case of occlusion of the object are respectively presented. As shown in the tracking results of **Fig. 10**, the traditional on-line MIL tracking method can cause the tracking to drift gradually when the object's attitude changes, and the tracking failure is known because of the accumulation of errors. The method of this paper combines the co-training algorithm, and effectively avoids the self-training problem of multiple instance learning algorithm and more accurate results are obtained.

Fig. 9. The proposed method

Fig. 10. MIL tracker

6. Conclusion

In this paper, we proposed a robot tracking system based on improved multiple instance learning and co-training. Firstly, aiming at the shortcoming of the lack of effective information in the sample obtained by using the Noisy-OR model in the multiple instance learning algorithm, based on the active learning method, we unify the sample selection and classifier training to selection more informative samples for the classifier training. Then, the co-training strategy is introduced into the algorithm, and the training is based on the classifier established by HOG and color histogram respectively. This makes the algorithm avoid drift and improve the tracking performance of the algorithm. Finally, we experiments on challenging video sequences and mobile robot demonstrated that the proposed algorithm performs well in terms of efficiency, accuracy and robustness.

Acknowledgment

This work is supported by Zhejiang Provincial Natural Science Foundation of China (No. LY18F030018, LZ15F020004), Natural Science Foundation of China (No.51376055, 61272311), 521 Plan of Zhejiang Sci-Tech University, and Science and Technology Plan of Zhejiang Province (No.2017C31017).

References

[1] Z. Pan, S. Liu, W. Fu, "A review of visual moving target tracking," *Multimedia Tools and Applications*, vol. 76, no. 16, pp. 16989-17018, 2017. Article (CrossRef LAink)

[2] S. Liu, Z. Pan, X. Cheng, "A novel fast fractal image compression method based on distance clustering in high dimensional sphere surface," *Fractals-Complex Geometry Patterns and Scaling in Nature and Society*, vol. 25, no. 4 , Article ID: 1740004, 2017. Article (CrossRef Link)

[3] Y.-D. Zhang, Y. Zhang, Y.-D. Lv, et al., "Alcoholism detection by medical robots based on Hu moment invariants and predator-prey adaptive-inertia chaotic particle swarm optimization," *Computers & Electrical Engineering*, vol. 63, pp. 126-138, 2017. Article (CrossRef Link)

[4] S. Liu, Z. Pan, H. Song, "Digital image watermarking method based on DCT and fractal encoding," *IET Image Processing*, vol. 11, no. 10, pp. 815-821, 2017. Article (CrossRef Link)

[5] W. Kim, J. Chun, "An improved approach for 3D hand pose estimation based on a single depth image and Haar random forest," *KSII Transactions on Internet and Information Systems*, vol. 9, no.8, pp. 3136-3150, 2015. Article (CrossRef Link)

[6] W. Choi, C. Pantofaru, S. Savarese, "A general framework for tracking multiple people from a moving camera," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, no. 7, pp.1577-1591, 2013. Article (CrossRef Link)

[7] R. Liu, Z. Du, L. Sun, "Moving object tracking based on mobile robot vision," in *Proc. of International Conference on Mechatronics and Automation*, pp.3625-3630, 2009. Article (CrossRef Link)

[8] S. Kim, J. Park, J. M. Lee, "Implementation of tracking and capturing a moving object using a mobile robot," *International Journal of Control Automation & Systems*, vol. 3, no. 3, pp. 444-452, 2005. Article (CrossRef Link)

[9] H. Lang, Y. Wang, W. D. S. Clarence, "Vision based object identification and tracking for mobile robot visual servo control," in *Proc. of IEEE International Conference on Control and Automation*, pp.92-96, 2010. Article (CrossRef Link)

[10] D. A. Ross, J. Lim, R. S. Lin, et al. "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1, pp. 125-141, 2008. Article (CrossRef Link)

[11] X. Mei, H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 33, no.11, pp. 2259-72, 2011. Article (CrossRef Link)

[12] K. Zhang, L .Zhang, M. H. Yang, "Real-time compressive tracking," in *Proc. of European Conference on Computer Vision*, pp.864-877, 2012. DOI : Article (CrossRef Link)

[13] S. Avidan, "Support vector tracking," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 26, no. 8, pp. 1064, 2004. Article (CrossRef Link)

[14] S. Avidan, "Ensemble tracking," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 29, no. 2, pp. 261-271, 2007. Article (CrossRef Link)

[15] R. T. Collins, Y. Liu, M. Leordeanu, "Online selection of discriminative tracking features," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 27, no. 10, pp.1631-1643, 2005. Article (CrossRef Link)

[16] H. Grabner, M. Grabner, H. Bischof, "Real-time tracking via on-line boosting," in *Proc. of British Machine Vision Conference 2006*, pp.47-56, 2006. Article (CrossRef Link)

[17] B. Babenko, M. H. Yang, S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp.1619-1632, 2011. Article (CrossRef Link)

[18] K. Zhang, H. Song, "Real-time visual tracking via online weighted multiple instance learning," *Pattern Recognition*, vol. 46, no.1, pp.397-411, 2013. Article (CrossRef Link)

[19] H. Grabner, C. Leistner, H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *Proc. of European conference on computer vision*, pp. 234-247, 2008. Article (CrossRef Link)

[20] K. Zhang, L. Zhang, Q. Liu, et al., "Fast visual tracking via dense spatio-temporal context learning," in *Proc. of European Conference on Computer Vision*, pp.127-141, 2014. Article (CrossRef Link)

[21] D. Wang, H. Lu, M. H. Yang, "Least soft-threshold squares tracking," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.2371-2378, 2013. Article (CrossRef Link)

[22] Y.-D.Zhang, Y. Zhang, X.-X.Hou, et al., "Seven-layer deep neural network based on sparse autoencoder for voxelwise detection of cerebral microbleed," *Multimedia Tools and Applications*, vol. 77, no. 9, pp. 10521-10538, 2018. Article (CrossRef Link)

[23] S.-H. Wang, Y.D. Lv, Y. Sui, et al., "Alcoholism detection by data augmentation and convolutional neural network with stochastic pooling," *Journal of Medical Systems*, vol. 42, no. 1, Article ID: 2, 2018. Article (CrossRef Link)

[24] B. Settles, "Active learning literature survey," *University of Wisconsin*, Madison, 2010. Article (CrossRef Link)

[25] T. M. Cover, J.A. Thomas, "Elements of information theory," *John Wiley & Sons*, 2005. Article (CrossRef Link)

[26] D. Zhang, F. Wang, Z. Shi, et al., "Interactive localized content based image retrieval with multiple-instance active learning," *Pattern Recognition*, vol. 43, no. 2, pp. 478-484, 2010. Article (CrossRef Link)

[27] F. Tang, S. Brennan, Q. Zhao, et al., "Co-tracking using semi-supervised support vector machines," in *Proc. of IEEE International Conference on Computer Vision*, pp.1-8, 2007. Article (CrossRef Link)

[28] Q. Yu, T. B. Dinh, G. Medioni, "Online tracking and reacquisition using co-trained generative and discriminative trackers," in *Proc. of European conference on computer vision*, pp. 678-691, 2008. Article (CrossRef Link)

[29] R. Liu, J. Cheng, H. Lu, "A robust boosting tracker with minimum error bound in a co-training framework," in *Proc. of IEEE International Conference on Computer Vision*, pp.1459-1466, 2009. Article (CrossRef Link)

[30] S. Liu, M. Lu, G.Liu, et al., "A novel distance metric: generalized relative entropy," *Entropy*, vol. 19, no. 6, Article ID: 269, 2017. Article (CrossRef Link)

[31] Y. Wu, J. Lim, M. H. Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834-1848, 2015. Article (CrossRef Link)

[32] Z. Kalal, K. Mikolajczyk, J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 34, no. 7, Article ID: 1409, 2012. Article (CrossRef Link)

[33] S. Hare, S. Golodetz, A. Saffari, et al., "Struck: structured output tracking with kernels," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 38, no.10, pp. 2096-2109, 2016. Article (CrossRef Link)

Zhiyu Zhou received M.S. and PhD. degree from Zhejiang Sci-Tech University in 2004 and 2018, respectively. He is now an associated professor at Zhejiang Sci-Tech University and mainly engaged in computer vision, grey systems and robotic tracking. He has published dozens of research papers, 10 papers among them have been included in SCI with the first author. He has just accomplished two Natural Science Fund Projects of Zhejiang Province, and now presided over another provincial Natural Science Fund Project.

Junjie Wang received M.S. degree from Zhejiang Sci-Tech University in 2018, he is mainly engaged in machine vision.

Yaming Wang received the Ph.D. degree in biomedical engineering from Zhejiang University, China. He is currently a Professor of computer science at Zhejiang Sci-Tech University, Zhejiang, China. He had been a Visiting Researcher and Visiting Scientist at Hong Kong University of Science and Technology, HKUST. His research interests include computer vision, pattern recognition and signal processing.

Zefei Zhu received his MS and PhD degrees from Zhejiang University in 1986 and 1999, respectively. Currently, he is a professor at Hangzhou Dianzi University and is mainly engaged in machine vision, mechanical automation, and manipulator control.

Jiayou Du received his MS degrees from Zhejiang University in 2006. Currently, he is a teacher at Hangzhou Dianzi University and is mainly engaged in machine vision, and microfluidics.

Xiangqi Liu received her MS degree from Zhejiang University in 2004 and PhD degree from Zhejiang Sci-Tech University in 2017.Currently, she is a lecturer at Hangzhou Dianzi University and is mainly engaged in mechanical engineering.

Jiaxin Quan received M.S. degree from Zhejiang Sci-Tech University in 2017, he is mainly engaged in machine vision.