Pattern Recognition 44 (2011) 2334–2342

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Boosted multi-class semi-supervised learning for human action recognition

Tianzhu Zhang [a,b,*], Si Liu [a,b], Changsheng Xu [a,b], Hanqing Lu [a,b]

[a] *National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
[b] *China-Singapore Institute of Digital Media, Singapore 119615, Singapore*

ARTICLE INFO

Available online 30 June 2010

Keywords:
Semi-supervised learning
Action recognition
adaBoost.MH
co-EM

ABSTRACT

Human action recognition is a challenging task due to significant intra-class variations, occlusion, and background clutter. Most of the existing work use the action models based on statistic learning algorithms for classification. To achieve good performance on recognition, a large amount of the labeled samples are therefore required to train the sophisticated action models. However, collecting labeled samples is labor-intensive. To tackle this problem, we propose a boosted multi-class semi-supervised learning algorithm in which the co-EM algorithm is adopted to leverage the information from unlabeled data. Three key issues are addressed in this paper. Firstly, we formulate the action recognition in a multi-class semi-supervised learning problem to deal with the insufficient labeled data and high computational expense. Secondly, boosted co-EM is employed for the semi-supervised model construction. To overcome the high dimensional feature space, weighted multiple discriminant analysis (WMDA) is used to project the features into low dimensional subspaces in which the Gaussian mixture models (GMM) are trained and boosting scheme is used to integrate the subspace models. Thirdly, we present the upper bound of the training error in multi-class framework, which is able to guide the novel classifier construction. In theory, the proposed solution is proved to minimize this upper error bound. Experimental results have shown good performance on public datasets.

1. Introduction

Human action recognition in video is receiving increasing attention due to its wide applications, such as content-based video retrieval, human–computer interfaces, video summarization, visual surveillance, etc. Human action recognition is a challenging research area because the dynamic human body motions have almost unlimited underlying representations. There also exist difficulties in perspective distortions, different viewpoints and illumination variations. Most of the existing work [1–3] stem from supervised learning scenario. To achieve good recognition performance, a large amount of labeled samples are needed in the training process [4–6]. However, labeled samples are usually difficult or expensive to obtain due to extensive labor cost. Therefore, how to achieve a good learning model with limited labeled samples is a crucial issue.

One way to reduce the amount of required labeled data is to develop algorithms that are able to learn from a small number of labeled examples augmented with a large number of unlabeled examples. Unlabeled examples, which can be easily obtained from public surveillance cameras, are much less expensive and easier to obtain than labeled examples. Recently, there has been interest in semi-supervised learning algorithms that utilize the labeled data as well as a large amount of unlabeled data to learn the hypothesis [7]. It shows great advantage by automatically exploiting huge amount of information from the unlabeled data and boosts the generalization ability of the trained hypothesis. To extract specific information from the unlabeled data, a number of semi-supervised learning methods, such as co-training and co-EM, are proposed [8–10].

Co-training [8] is a semi-supervised, multi-view algorithm that uses the initial training set to learn a (weak) classifier in each view. Then the learned classifiers are used to label all unlabeled examples, and find out those examples whose labels are most confident by classifiers. These high-confidence examples are labeled with the estimated class labels and added to the training set. Based on the new training set, a new classifier is learnt in each view, and the whole process is repeated for several iterations. At the end, a final hypothesis is created by a voting scheme that combines the prediction of the classifiers learnt in each view. To use unlabeled data, some works combine boosting and co-training to construct learning approach [11–13], which are efficient to exploit the unlabeled data.

Compared with co-training, co-EM algorithm [9,10] can be thought of as a closer match to the theoretical argument of Blum and Mitchell [8]. Moreover, co-EM algorithm does not commit to a label for the unlabeled examples; instead, it uses probabilistic

[*] Corresponding author at: National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
Tel.: +86 138 1061 4386.
E-mail address: tzzhang@nlpr.ia.ac.cn (T. Zhang).

0031-3203/$ - see front matter Crown Copyright © 2010 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.06.018

labels that may change from one iteration to the other. By contrast, co-training's commitment to the high-confidence predictions may add to the training set a large number of mislabeled examples, especially during the first iterations, when the hypotheses may have little prediction power. In addition, co-EM converges as quickly as EM does. Despite its popularity and usefulness, co-EM algorithm suffers from insufficient training data, especially when the feature space is of high dimensionality. This restricts the applicability of co-EM to situations where there are plenty of training data.

For the human action recognition task, there are many scenarios of multiple labels. Therefore it will be useful to generalize an algorithm to the multi-class form. Several extensions of adaBoost for multi-class problems have been suggested [14,15]. In this work we extend the adaBoost.MH [15] algorithm to co-EM case. By combination of the adaBoost.MH and co-EM, we propose a novel boosted multi-class semi-supervised learning algorithm for human action recognition. In our approach, the data are described as a finite hierarchical Gaussian mixture model (GMM). To avoid the insufficient training data, especially when the feature space is of high dimensionality, a weighted multiple discriminant analysis (WMDA) is adopted to make the required amount of training data depending only upon the number of classes, regardless of the feature dimension. Then the co-EM algorithm is employed to learn the GMM in the WMDA subspace by probabilistically labeling all unlabeled examples and iteratively exchanging those labels between two views (features). Consequently, a set of weak hypotheses for each view are learnt in the boosting framework and finally a strong classifier is obtained for action recognition. For the proposed algorithm, a derived boosting error bound is served as the theoretical guidance for the training error.

The proof of our work is similar in spirit to Liu et al. [13]'s efforts to combine adaboost and co-training for tracking. The key difference is that we focus on developing a boosted multi-class semi-supervised learning algorithm for action recognition with the co-EM algorithm. Most of the existing action classification algorithms are based on one-against-all strategy, in which each action category is trained with a classifier. Compared with the extensively used one-against-all classification strategy, a multi-class recognition algorithm only needs to train one unified model which is less computation-intensive.

Compared with the previous methods, our algorithm has the following advantages:

- A boosted multi-class semi-supervised learning algorithm is proposed for human action recognition, which is efficient to combine labeled and unlabeled samples to improve the recognition performance.
- A WMDA is adopted to make the co-EM algorithm efficiently learn parameters regardless of the feature dimension and avoid re-sampling the training data. In addition, boosting the GMM in a series of WMDA subspaces enhances the discriminative power of our algorithm.
- For this boosted multi-class semi-supervised learning algorithm, a derived upper error bound is served as the theoretical guidance for classifier construction.

2. Related work

In this section, we mainly focus on existing methods related to our work. Boosting and co-EM are two key components of our approach for action recognition. We briefly review the work related to action recognition and the error analysis for adaboost.MH and co-EM.

2.1. Action recognition

To represent human actions, some significant efforts have been made in spatio-temporal volumes [16,17], spatio-temporal interest points [18,19,1,5] or trajectory [20,21]. Recently, some approaches use the combination of appearance and motion features [22,5,2]. Laptev et al.'s spatio-temporal interest points [1] have shown good performance for action recognition and are adopted in this paper. Histograms of oriented gradient (HoG) and optical flow (HoF) are considered as two "views", in the co-EM algorithm. To recognize human action, a lot of works use labeled samples to train action models [21,23]. Alternatively, some researchers work on directly learning from unlabeled action dataset in a unsupervised manner [24–26]. However, there are very few semi-supervised learning methods for human action analysis, which can fully use both the labeled and unlabeled data. Guan et al. [27] propose an En-co-training method to make use of the unlabeled action videos. It shows that the learning performance can be improved by utilizing the unlabeled data, but the comparative experimental results with the state of the art methods on publicly dataset are not reported.

2.2. Hamming loss of adaBoost.MH

In [15], Schapire and Singer show that the following bound holds for the Hamming loss of H on the training data:

$$hloss(H) = \frac{1}{nL}\sum_{i,l}[\![\operatorname{sign}(H(x_i,l)) \neq Y_i[l]]\!] \leq \prod_{t=1}^{T} Z_t, \qquad (1)$$

where x_i is the ith training sample and $Y_i[l]$ is the corresponding class label. For any predicate $*$, let $[\![*]\!]$ be 1 if $*$ holds and 0 otherwise. n is the number of the training samples and Z_t is a normalization factor which is the weight sum of all the samples after the tth weak hypothesis training. Through minimizing Z_t in each weak hypothesis learning, adaBoost.MH decreases the whole error upper bound of itself. The Z_t can be expressed by

$$Z_t = \sum_{i,l} D_t(i,l)\exp(-\alpha_t Y_i[l] h_t(x_i,l)), \qquad (2)$$

where $D_t(i,l)$ is the normalized weight of the ith sample in the tth weak hypothesis training.

2.3. Upper error bound of co-EM

Dasgupta et al. [28] give PAC bounds on the error of co-training in terms of the disagreement rate of hypotheses on unlabeled data in two independent views. This justifies the direct minimization of the disagreement. Our analysis is mainly based on the work in [29]. It proves that PAC-style guarantees that if two independent hypotheses $h^j(x)$ in views $j=1$, 2 have a probability at least 50% of assigning x to the correct label, then with high probability the misclassification rate is upper bounded by the rate of disagreement between the classifiers based on each view. The above error bound can be approximately expressed as follows:

$$P(h^1(x) \neq h^2(x)) \geq \max_j P(h^j(x) \neq y), \qquad (3)$$

where y is the real label, $j \in \{1,2\}$ is the index of the view and $h^j(x)$ is the classifier based on the jth view. In unsupervised learning, the risk of assigning instances to wrong labels cannot be minimized directly, but this argument shows that we can minimize an upper bound on this risk by minimizing the disagreement of multiple hypotheses.

3. Boosted multi-class semi-supervised learning algorithm

Based on the adaBoost.MH and co-EM, we propose a novel boosted multi-class semi-supervised learning algorithm. The proposed algorithm trains two classifiers in different views combining labeled and unlabeled samples using the co-EM algorithm.

In the multi-label case, each instance $x \in \mathcal{X}$ may belong to multiple labels in $\mathcal{Y}$, where $\mathcal{Y}$ is a finite set of labels, and let $L = |\mathcal{Y}|$. Thus, a labeled example is a pair (x, Y) where $Y \subseteq \mathcal{Y}$ is the set of labels assigned to x. The single-label case is clearly a special case in which $|Y| = 1$ for all observations.

For $Y \subseteq \mathcal{Y}$, define $Y[l]$ for $l \in \mathcal{Y}$ to be

$$Y[l] = \begin{cases} +1 & \text{if } l \in Y, \\ -1 & \text{if } l \notin Y. \end{cases}$$

The classifier of adaBoost.MH is an ensemble of several weak hypotheses:

$$H(x,l) = \sum_{t=1}^{T} \alpha_t h_t(x,l), \tag{4}$$

where $H(x,l)$ is the ensemble strong hypothesis and its classification result is $\text{sign}(H(x,l))$, $h_t(x,l)$ is the tth weak hypothesis to be learned and α_t is the corresponding voting weight.

To treat unlabeled data in the updated process, two views are adopted to describe each sample in the co-EM framework. Classifiers based on each view are trained iteratively, and then naturally combined together to give a final hypothesis. Such multi-view ensemble classifier can be represented as

$$F(x,l) = \sum_j H^j(x,l), \tag{5}$$

where $H^j(x,l)$ is the strong classifier based on the jth view defined in Eq. (4). In Sections 3.1 and 3.2, we will describe the detail of the boosted multi-class semi-supervised learning algorithm, following the key issues introduced in [15].

3.1. Designing weak hypothesis h_t

Hierarchical mixture models provide flexible discrimination tools, where each conditional distribution $f^j(x|y=l)$ in view j is modelled by a mixture of components [30]. At the high level, the distribution is described by

$$f^j(x|\phi^j) = \sum_{k=1}^{K} w_k^j f_k^j(x;\theta_k),$$

where K is the number of components, w_k the mixing proportions, θ_k the conditional distribution parameters, and ϕ^j denotes all parameters $\{w_k^j;\theta_k\}_{k=1}^K$. The high-level description can also be represented as a low-level mixture of components, as shown here for multi-class classification:

$$f^j(x|\phi^j) = \sum_{l=1}^{L}\sum_{k_l=1}^{K_l} w_{k_l}^j f_{k_l}^j(x;\theta_{k_l}). \tag{6}$$

In our hierarchical mixtures, the class label y provides a subset of possible components. When $y = l$ the modes from k_l=1 to K_l are possible, and when an example is unlabeled, all modes are possible. In this setting, the co-EM algorithm can be used to maximize the log-likelihood with respect to ϕ^j to obtain the parameters with the unlabeled data.

Once the parameters are learnt, the probability of a sample x belonging to each class l can be obtained as

$$p_l^j(l|x) = \sum_{k_l=1}^{K_l} w_{k_l}^j f_{k_l}^j(x;\theta_{k_l}). \tag{7}$$

Based on these probabilities, we can design weak hypothesis as follows:

$$h^j(x,l) = \begin{cases} +1 & \text{if } p_l^j \geq th_l^j, \\ -1 & \text{otherwise}, \end{cases} \tag{8}$$

where $p_l^j = p_l^j(l|x)/\sum_{l=1}^{L} p_l^j(l|x)$, th_l^j is a threshold determined by minimizing the weighted classification error.

The Gaussian mixture model (GMM) can approximate arbitrary probability distributions, which makes it a powerful tool for feature representation and classification. However, it suffers from insufficient training data using EM or co-EM to learn its parameters, especially when the feature space is of high dimensionality. To tackle this problem, Tang and Huang [31] proposed a method to boost the GMMs via discriminant analysis. At each iteration t, a multiple discriminant analysis (MDA) subspace is created by projection of the re-sampled training examples according to a distribution $D_t(i)$, which is adaptively adjusted to assign more weights to those examples misclassified by the GMM classifier in the previous iteration. However, the re-sampling process does not make use of sample weights and leads inefficient use of the training data. To overcome this problem, we train GMM in a weighted multiple discriminant analysis (WMDA) subspace using co-EM algorithm. The data are projected from the high dimensional space to a $(L-1)$-dimensional subspace, in which the required amount of training data depends only upon the number of classes, regardless of the feature dimension. The weighted mean and scatter matrices are defined as

$$m_l = \frac{\sum_i^{n_l} D_t(i)x_i}{\sum_i^{n_l} D_t(i)}, \quad S_W = \sum_{l=1}^{L} \frac{\sum_i^{n_l} D_t(i)(x_i - m_l)(x_i - m_l)^T}{\sum_i^{n_l} D_t(i)},$$

$$m = \frac{\sum_i^{n} D_t(i)x_i}{\sum_i^{n} D_t(i)}, \quad S_B = \sum_{l=1}^{L}\sum_i^{n_l} \frac{D_t(i)(m_l - m)(m_l - m)^T}{\sum_i^{n_l} D_t(i)}, \tag{9}$$

where

$$D_t(i) = \sum_{l=1}^{L} D_t(i,l), \tag{10}$$

n_l is the number of samples in class l, n is the number of all samples, and $D_t(i,l)$ are the weights provided by boosting at step t.

Using WMDA can eliminate the need of re-sampling the training data. This in turn leads to a more efficient use of the training data. The GMM formulated in the WMDA subspace requires significantly less training data than that in the original high-dimensional space. Although important discriminatory information may be lost during dimensionality reduction, we believe that the loss of discriminative power due to dimensionality reduction can be effectively compensated through iteratively boosting the GMM in a series of different WMDA subspaces. By way of boosting the different GMM, the samples are classified into different low-dimensional subspaces. Finally, boosting scheme is adopted to integrate the different subspaces models, and experimental results validate the effectiveness of our proposed method.

3.2. Choosing α_t

In semi-supervised learning, if the labels of all samples are given as in Eq. (11), in order to minimize training error, a reasonable approach might be to greedily minimize Z_t on each round of boosting for multiple views. In this work, our research is restricted to two views, $j=1,\ 2$. This will be proved in Section 4. We can apply this idea in the choice of α_t in Eq. (4):

$$u_i[l] = \begin{cases} Y_i[\mathrm{l}], & i=1,\ldots,m, \\ \mathrm{sign}(h_t^{3-j}(x_i,l)), & i=m+1,\ldots.,n \end{cases} \tag{11}$$

For fixed t, let $W_{t,+}^j$, $W_{t,-}^j$ be defined by

$$W_{t,+}^j = \sum_{i,l:h_t^j(x_i,l)=u_i[l]} D_t^j(i,l),$$

$$W_{t,-}^j = \sum_{i,l:h_t^j(x_i,l)\neq u_i[l]} D_t^j(i,l). \tag{12}$$

We can calculate Z_t^j as

$$\begin{aligned} Z_t^j &= \sum_{i,l} D_t^j(i,l)\exp(-u_i[l]\alpha_t^j h_t^j(x_i,l)) \\ &= W_{t,+}^j e^{-\alpha} + W_{t,-}^j e^{\alpha}. \end{aligned} \tag{13}$$

It can be easily verified that Z_t^j is minimized when

$$\alpha_t^j = \frac{1}{2}\ln\left(\frac{W_{t,+}^j}{W_{t,-}^j}\right). \tag{14}$$

For this setting of α_t^j, we have

$$Z_t^j = 2\sqrt{W_{t,+}^j W_{t,-}^j}.$$

The weight updates of the samples are obtained as follows:

$$D_{t+1}^j(i,l) = \frac{D_t^j(i,l)\cdot\exp(-Y_i[l]\alpha_t^j h_t^j(x_i,l))}{Z_t}, \quad i=1,\ldots,m,$$

$$D_{t+1}^j(i,l) = \frac{D_t^j(i,l)\cdot\exp(-sign(h_t^{3-j}(x_i,l))\alpha_t^j h_t^j(x_i,l))}{Z_t}, \quad i=m+1,\ldots,,n. \tag{15}$$

The final strong hypothesis described in Eq. (5) is obtained by linear combination of all the weak hypotheses in each view, i.e. $F(x,l)=\sum_{j=1}^{2}\sum_t \alpha_t^j h_t^j(x,l)$.

A more detailed description of the proposed boosted multi-class semi-supervised learning algorithm is given in Algorithm 1. In each iteration, a WMDA subspace is created by projection of the labeled training examples according to a distribution $D_t(i)$ as shown in Eq. (10). Based on the WMDA subspace, GMM is learned by using co-EM algorithm combining labeled and unlabeled samples. The co-EM algorithm probabilistically labels all unlabeled examples and iteratively exchanges those labels between two views. This process iterates until the classifiers converge, and the number of iterations is about 25. Once the $p_l^j(l|x)$ is learnt, the weak hypothesi $h_{t,0}^j$ for each view is initialized just using the labeled data. The unlabeled data can be set with pseudo-labels and the hypothesis is obtained with the lowest training error as shown in Eq. (8), and this process is repeated for $\mathcal{K}$ times. The weak hypothesis with the lowest error is selected, and the voting weight is calculated. Then the sample weights are updated. After T iterations, a final strong hypothesis can be learned. In Algorithm 1, $h_{t,k}^j(x,l)$ is the weak hypothesis selected in the kth iteration for the jth view. u_i is the label of the ith sample. For the labeled sample, u_i is the real label. For the unlabeled sample, u_i is the pseudo-label labeled by weak hypothesis in different view.

Algorithm 1. The proposed boosted multi-class semi-supervised learning algorithm

```
Input: labeled examples $X_\mathbf{l}=\{\langle x_i,Y_i\rangle | i=1,\ldots,m\}$ and
unlabeled examples $X_\mathbf{u}=\{\langle x_i,\bullet\rangle | i=m+1,\ldots,n\}$,
$X=X_\mathbf{l}\cup X_\mathbf{u}$.
Initialize: $\forall i,j: D_1^j(i,l)=\frac{1}{nL}, j=1,2;\ l=1,\ldots,L$
for $t=1,\ldots,T$ : do
    Project $X_\mathbf{l}$ to WMDA subspace according to the
    distribution $D_t(i)$ shown in Eq. (10).
    Train GMM in the WMDA subspace for $x\in X$ using
    co-EM algorithm to obtain $p_l^j(l|x)$ in each view.
    initialize $\forall j: h_{t,0}^j$ using only labeled data $X_\mathbf{l}$ in the
    WMDA subspace according to the Eq. (8).
    for $k=1,\ldots,\mathcal{K}$ and $j=1,2$ do
        Set pseudo-labels:
        $u_i[l]=\begin{cases} Y_i[l] & i=1,\ldots,m \\ \mathrm{sign}(h_{t,k-1}^{3-j}(x_i,l)) & i=m+1,\ldots,n \end{cases}$
        Choose hypothesis $h_{t,k\text{-}1}^j$ with the lowest error as in
        Eq. (8) and get
        $W_{t,-,k,l}^j=\sum_{i:h_{t,k}^j(x_i,l)\neq u_i[l]} D_t^j(i,l)$
    end for
    Output selected hypothesis $h_t^j(x,l)$:
    $h_t^j(x,l)=h_{t,k_0}^j(x,l)$, where $k_0=\arg\min_{k\in\{1,\ldots,\mathcal{K}\}} W_{t,-,k,l}^j$.
    Calculate voting weight with Eq. (12):
    $\alpha_t^j=\frac{1}{2}\ln\left(\frac{W_{t,+}^j}{W_{t,-}^j}\right)$
    Update sample weights for $\forall i,j$:
    $D_{t+1}^j(i,l)=\frac{D_t^j(i,l)\exp(-\alpha_t^j u_i[l]h_t^j(x_i,l))}{Z_t^j}$
    where $Z_t^j=\sum_{i,l} D_t^j(i,l)\exp(-\alpha_t^j u_i[l]h_t^j(x_i,l))$
end for
Output: the final strong hypothesis
$F(x,l)=\sum_{j=1}^{2}\sum_{t=1}^{T}\alpha_t^j h_t^j(x,l)$
```

4. Error analysis of the proposed algorithm

In this section, we discuss the hamming loss of our boosted multi-class semi-supervised algorithm and give a simple justification in theory.

For the jth view, where $j\in\{1,2\}$, we have the following normalized weight of the ith sample:

$$D_{T+1}^j(i,l) = \frac{\exp(-Y_i[l]H^j(x_i,l))}{nL\prod_{t=1}^{T} Z_t^j}, \tag{16}$$

where $D_1^j(i,l)=1/nL$. If $\mathrm{sign}(F(x_i,l))\neq Y_i[l]$ then $Y_i[l]F(x_i,l)\leq 0$ implying that $\exp(-Y_i[l]F(x_i,l))\geq 1$ and $[\exp(-Y_i[l]F(x_i,l))]^{1/J}\geq 1$. Thus,

$$\begin{aligned} [\![\, \mathrm{sign}(F(x_i,l))\neq Y_i[l] \,]\!] &\leq [\exp(-Y_i[l]F(x_i,l))]^{1/J} \\ &= \left[\exp\left(-Y_i[l]\sum_{j=1}^{J} H^j(x,l)\right)\right]^{1/J} = \prod_{j=1}^{J}[\exp(-Y_i[l]H^j(x_i,l))]^{1/J} \\ &\leq \frac{1}{J}\sum_{j=1}^{J}\exp(-Y_i[l]H^j(x_i,l)). \end{aligned} \tag{17}$$

Combining Eqs. (16) and (17), the following bound holds on the training error of $F(x,l)$ in supervised leaning:

$$\begin{aligned} hloss(F) &= \frac{1}{nL}\sum_{i,l}[\![\, \mathrm{sign}(F(x_i,l)) \neq Y_i[l]\,]\!] \\ &\leq \frac{1}{nL}\sum_{i,l}\left[\frac{1}{J}\sum_{j=1}^{J}\exp(-Y_i[l]H^j(x_i,l))\right] \\ &= \frac{1}{nL}\sum_{i,l}\left\{\frac{1}{J}\sum_{j=1}^{J} nL\left(\prod_{t=1}^{T} Z_t^j\right)D_{T+1}^j(i,l)\right\} \\ &= \frac{1}{J}\sum_{j=1}^{J}\left[\left(\prod_{t=1}^{T} Z_t^j\right)\sum_{i,l}D_{T+1}^j(i,l)\right] = \frac{1}{J}\sum_{j=1}^{J}\left(\prod_{t=1}^{T} Z_t^j\right). \end{aligned} \tag{18}$$

In semi-supervised learning, the error bound of $F(x,l)$ is derived as follows.

Theorem 1. *Assuming the upper error bound of co-EM comes true in Eq.* (3), *the following bound holds on the training error of* $F(x,l)$ *in semi-supervised leaning*:

$$\begin{aligned} hloss(F) &= \frac{1}{nL}\sum_{i,l}[\![\, \mathrm{sign}(F(x_i,l)) \neq Y_i[l]\,]\!] \\ &\leq \frac{1}{2nL}\left\{\sum_{i=1}^{m}\sum_{l=1}^{L}\exp(-Y_i[l]H^1(x_i,l))\right. \\ &\quad + \sum_{i=1}^{m}\sum_{l=1}^{L}\exp(-Y_i[l]H^2(x_i,l)) \\ &\quad + \sum_{i=m+1}^{n}\sum_{l=1}^{L}\exp(-\mathrm{sign}(h_t^2(x_i,l))\sum_{t=1}^{T}\alpha_t^1 h_t^1(x_i,l)) \\ &\quad \left. + \sum_{i=m+1}^{n}\sum_{l=1}^{L}\exp(-\mathrm{sign}(h_t^1(x_i,l))\sum_{t=1}^{T}\alpha_t^2 h_t^2(x_i,l))\right\}. \end{aligned}$$

The proof of this Theorem 1 appears in Appendix A. The important consequence of Theorem 1 is that, in semi-supervised learning, if the labels of all samples are given as in Eq. (11), in order to minimize training error, a reasonable approach might be to greedily minimize the bound given in Eq. (18) by minimizing Z_t^1 and Z_t^2 on each round of boosting. We apply this idea both in the choice of α_t and as a general criterion for the choice of weak hypothesis h_t introduced in Section 3.

5. Experimental results

In this section, we systematically evaluate the effectiveness of our proposed boosted multi-class semi-supervised learning algorithm on public action recognition datasets, i.e. the HOHA database of movie videos used in [1] and the KTH actions dataset [19].

5.1. Experimental settings

The HOHA database contains two video training sets, a manual and an automatic one, and a video test set. It contains video clips for eight classes of movie actions. In our work, the manual training set and the test set are used, and they contain 430 videos with clean label. This database is collected from realistic movie and is very complicated. The KTH database contains 600 low-resolution (160 × 120) video files, and each file contains four sequences. There are about 2391 sequences of 25 people, each performing six natural actions: 'boxing', 'handclapping', 'handwaving', 'jogging', 'running' and 'walking'. Each action is performed under four different conditions: outdoors, outdoors with scale variations, outdoors with different clothes and indoors. Each video contains one person repeatedly performing one action for four times. The challenges in this database include scale changes, action frequency changes and illumination variations.

We use Laptev's spatio-temporal interest point features to describe video [1]. The currently implemented types of descriptors are HOG (histograms of oriented gradients) and HOF (histograms of optical flow) computed on a 3D video patch in the neighborhood of each detected STIP. The patch is partitioned into a grid with $3\times3\times2$ spatio-temporal blocks; 4-bin HOG descriptors and 5-bin HOF descriptors are then computed for all blocks and are concatenated into a 72-element and 90-element descriptors, respectively. The details can be found in [1]. In our work, these HOG and HOF features are considered as the $View_1$ and $View_2$, respectively. Due to the limitations of the distributed implementation of Laptev's spatio-temporal interest point features, we do not adopt the best performing channel combination. On the KTH dataset, we use k-means to get a visual vocabulary, the number of clusters is empirically set to 1000. On the HOHA dataset, the number of clusters is set to 4000.

HOHA and KTH datasets contain single-action per video sequence, and most of the previously published results assign a single label to each sequence (per video sequence classification). As a result, we also report per video sequence classification on these two datasets. Note that the testing and training samples on KTH dataset are video sequences obtained according to the sequence boundaries given by 'www.nada.kth.se/cvap/actions/00sequences.txt'.

5.2. Comparison on HOHA dataset

On the HOHA database, we randomly choose $\frac{1}{2}$ of the whole dataset as training samples T. We again randomly select $\frac{1}{4}$ as additional training samples A. The remaining $\frac{1}{4}$ data are left for testing. The results are shown in Table 1. In method 1, we combine T and A as the enlarged training set. We ignore the label information of A and combine the unlabeled A with T to form the training set in method 2. Different from the first two experiment settings, we only make use of T in method 3. The average accuracy are 36.23%, 32.88%, 28.53% respectively. From Table 1, the accuracy of method 1 is lower than method 2 for AnswerPhone, and StandUp. The reason is some unlabeled samples from other classes are very similar to the training data of the two classes, and it leads to the false classification in test dataset. To fairly compare with [1], we follow the same experiment setting and use the same features: HoG-BoF and HoF-BoF as two views. Our average

Table 1
The comparison of different methods about average precision (AP) for each action class on the HOHA dataset.

Action	Method		
	1 (%)	2 (%)	3 (%)
AnswerPhone	33.68	38.14	31.91
GetOutCar	38.88	32.95	31.50
HandShake	33.50	27.27	17.42
HugPerson	36.75	35.56	26.46
Kiss	54.75	45.35	40.56
SitDown	29.81	21.50	25.30
SitUp	11.63	10.37	9.45
StandUp	50.88	51.93	45.65

accuracy is 30.51% on both views which is better than 27.00% using HoG-BoF and 21.49% using HoF-BoF, respectively. In a single view, the average accuracy is 29.83% and 26.79% using HoG-BoF and HoF-BoF respectively. This result does not exceed the 38.39% which is reported by using the best channel combination. It shows that the feature is also important in action recognition. How to describe action efficiently is another problem beyond the discussion scope of this paper. For our method, we can draw the conclusion that it is useful to exploit unlabeled samples to improve performance.

In addition, we randomly select $\frac{3}{8}$ of the whole dataset as training samples T and choose $\frac{1}{8}$ as additional training samples A. The remaining $\frac{4}{8}$ samples are used for testing. We do three experiments that are similar to the above three methods, and the average accuracies are 27.16%, 25.62%, 22.83% respectively. From these results, we can come to the conclusion that our method is effective to adopt unlabeled samples to improve performance and alleviate the human labors to manually label training samples.

5.3. Comparison on KTH dataset

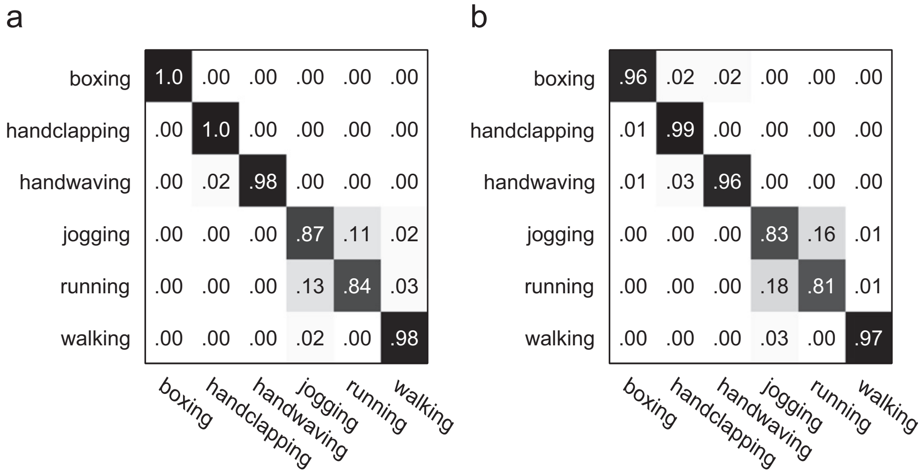
In order to evaluate our proposed algorithm, we perform experiments with three different configuration on the KTH dataset. For the first method, we adopt the leave-one-out cross validation scheme and use supervised learning. Details are as follows. Twenty-four out of the 25 actors in the database are used to train the classifier, the 25th is used for the evaluation. This is repeated for all 25 actors and the rates are averaged. The confusion matrix is given in Fig. 1(a). The average accuracy is about 94.5%. For the second method, we use 15 out of the 25 actors as labeled samples, nine actors as unlabeled samples, and the 25th is used for testing. For each actor, there are about 96 video sequences. This is repeated for all 25 actors and the rates are averaged. The confusion matrix for this method is given in Fig. 1(b). The average accuracy is 92.0%. The third method is mostly the same as the second. The only difference is that we do not use the unlabeled data. The accuracy is 88.1%. As the dataset is small, the number of Gaussian is set to 3 for each class l in our experiments.

Table 2 compares the average class accuracy of our first method with results reported by other researchers. Compared with the existing approaches, our method shows much better performance in supervised learning, outperforming the state-of-the-art approaches. Even in semi-supervised learning where only small labeled samples are used, our second method achieves a good performance. Moreover, the result of our second method is better than our third method, which shows that our semi-supervised learning can make use of unlabeled samples to improve the performance. Note that a precise comparison between the approaches is difficult, since experimental setups, e.g. different strategy in training, slightly differ with each approach.

In addition, we realize the first method on the KTH dataset with one-against-all classification strategy and our multi-class recognition algorithm to compare the computation time. For our multi-class recognition algorithm, it costs about 9.6 min. For the one-against-all classification strategy, the computation time is about 21.5 min. The results validate our method is less computation-intensive than the one-against-all classification strategy. The maximum numbers of iterations in the boosting process and in the co-EM process are manually set to 80 and 35, respectively. All the experiments are run on a sever with Intel(R) Pentium(R) Dual CPU with 2.2 GHz CPU and 2 GB memory and the code is run in MATLAB platform.

However, only using small unlabeled samples cannot evaluate the performance of our algorithm efficiently, hence we attempt to use more unlabeled samples to test the proposed method. On KTH dataset, each video contains one person repeatedly performing one action, therefore we divide each video into multiple short sequences with $\frac{1}{2}$ overlap. In this way, we obtain about 6295 short videos. We prevent the same subject from doing the same action to be in both training and testing portions. In these videos, 960 videos are used for testing, and the remaining are used as labeled and unlabeled data. We evaluate our proposed algorithm on various size of labeled and unlabeled data. As Fig. 2(a) and (b) shows, different size of labeled and unlabeled data are chosen, respectively. The performance is obviously improved when 2500 unlabeled data are adopted compared with that of without unlabeled data. It can be seen that with more unlabeled data for training, the system gets better performance. It also shows that unlabeled data help to improve the performance of the system.

We also do experiments to compare our proposed method with boosted co-training method for multi-class classification. This method is different from Algorithm 1 at the 4th step. The classifiers in two views are initialized using EM algorithm with labeled samples for each class l. At every round of co-training each classifier labels and chooses unlabeled samples per class to be added to the labeled samples. The comparison results are shown in Fig. 3. As can be seen, our proposed method outperforms the boosted co-training algorithm, especially at the previous iterations.



a

	boxing	handclapping	handwaving	jogging	running	walking
boxing	1.0	.00	.00	.00	.00	.00
handclapping	.00	1.0	.00	.00	.00	.00
handwaving	.00	.02	.98	.00	.00	.00
jogging	.00	.00	.00	.87	.11	.02
running	.00	.00	.00	.13	.84	.03
walking	.00	.00	.00	.02	.00	.98

b

	boxing	handclapping	handwaving	jogging	running	walking
boxing	.96	.02	.02	.00	.00	.00
handclapping	.01	.99	.00	.00	.00	.00
handwaving	.01	.03	.96	.00	.00	.00
jogging	.00	.00	.00	.83	.16	.01
running	.00	.00	.00	.18	.81	.01
walking	.00	.00	.00	.03	.00	.97

Fig. 1. Confusion matrices on the KTH dataset. (a) Results using supervised learning strategy. (b) Results using semi-supervised learning strategy.

Table 2
The comparison of different methods on the KTH dataset.

Method	Schuldt et al. [19]	Niebles et al. [24]	Laptev et al. [1]	Liu et al. [2]	Ours
Accuracy (%)	71.7	83.3	91.8	93.8	94.5

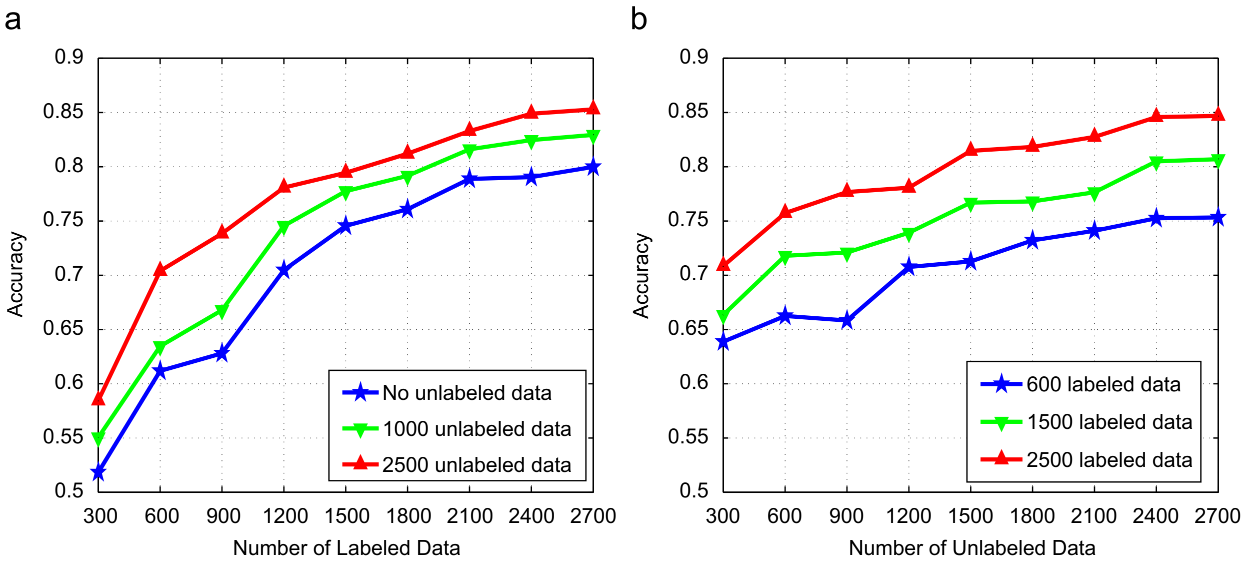


Fig. 2. Evaluation the performance of the proposed method with different number of unlabeled data and labeled data. (a) Comparison of accuracy on various number of labeled data. (b) Comparison of accuracy on various number of unlabeled data.

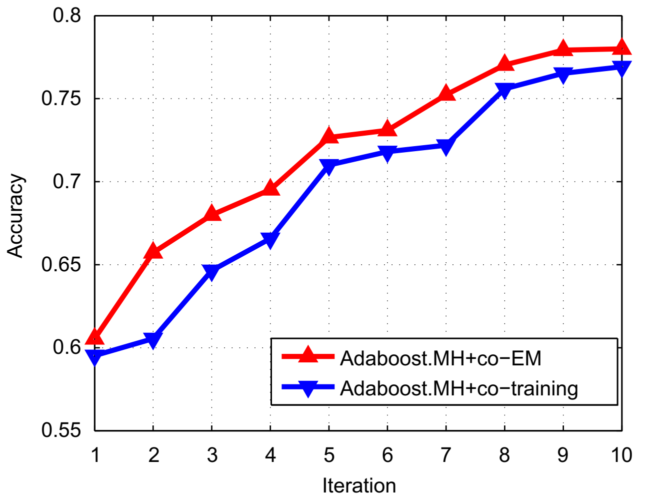


Fig. 3. Comparison between boosted co-EM and boosted co-training on KTH dataset.

6. Conclusions

We have proposed a boosted multi-class semi-supervised learning algorithm for human action recognition by combining adaboost.MH and co-EM. Through the co-EM algorithm, labeled data and unlabeled data are adopted to train classifiers in two views. To avoid suffering from insufficient training data, especially when the dimension of the feature space is high, a WMDA is employed to make use of training data and learn the parameters of GMM by co-EM efficiently. In addition, we give a theoretical analysis of the training error of our algorithm. We have tested our proposed method on public human action databases, the results are encouraging. We believe that the proposed boosted multi-class semi-supervised learning algorithm can be also helpful for other applications where manual labels are costly. To improve the performance of action recognition, we will explore more effective representations for action information in the future.

Acknowledgements

The research is supported by the following funding resources: NSFC (National Natural Science Foundation of China) #60970092, #60975010 and #60835002; 973 Program Project #2010CB327905.

Appendix A

This appendix gives a simple proof of Theorem 1.

Proof. In semi-supervised learning, $\{x_i | i = 1, \ldots, m\}$ are the labeled samples and $\{x_i | i = m+1, \ldots, n\}$ are the unlabeled samples. The hamming loss of Theorem 1 can be represented as follows by Eq. (18):

$$\begin{aligned} hloss(F) &= \frac{1}{nL}\sum_{i,l}[\![\operatorname{sign}(F(x_i,l)) \neq Y_i[l]]\!] \\ &\le \frac{1}{2nL}\left\{ \sum_{i=1}^{m}\sum_{l=1}^{L}\exp(-Y_i[l]H^1(x_i,l)) + \sum_{i=1}^{m}\sum_{l=1}^{L}\exp(-Y_i[l]H^2(x_i,l)) \right. \\ &\quad \left. + \sum_{i=m+1}^{n}\sum_{l=1}^{L}\exp(-Y_i[l]H^1(x_i,l)) + \sum_{i=m+1}^{n}\sum_{l=1}^{L}\exp(-Y_i[l]H^2(x_i,l)) \right\}. \end{aligned} \tag{19}$$

In Eq. (19), the first and second terms have upper error bounds as shown in Eq. (18), however, $\{Y_i[l] | i = m+1, \ldots, n\}$ is unknown in the third and fourth terms. Combining Eqs. (2) and (16), the third term on the right side of Eq. (19) can be transformed as follows:

$$\begin{aligned}\sum_{i=m+1}^{n}\sum_{l=1}^{L}\exp(-Y_i[l]H^1(x_i,l)) &= \sum_{i=m+1}^{n}\sum_{l=1}^{L}\left[nL\left(\prod_{t=1}^{T}Z_t^1\right)D_{T+1}^1(i,l)\right]\\ &= nL\sum_{i=m+1}^{n}\sum_{l=1}^{L}\left[\left(\prod_{t=1}^{T}Z_t^1\right)D_{T+1}^1(i,l)\right]\\ &= nL\sum_{i=m+1}^{n}\sum_{l=1}^{L}(D_{T+1}^1(i,l))\prod_{t=1}^{T}Z_t^1\\ &= nLD\prod_{t=1}^{T}\left(\sum_{i,l}D_t^1(i,l)\exp(-\alpha_t^1 Y_i[l]h_t^1(x_i,l))\right)\\ &= nLD\prod_{t=1}^{T}[W_{t,+}^1\cdot\exp(-\alpha_t^1)+W_{t,-}^1\cdot\exp(\alpha_t^1)]\\ &= nLD\prod_{t=1}^{T}[\exp(-\alpha_t^1)+W_{t,-}^1\cdot(\exp(\alpha_t^1)-\exp(-\alpha_t^1))], \qquad (20)\end{aligned}$$

where $W_{t,+}^1=\sum_{i,l:Y_i[l]=h_t^1(x_i,l)}D_t^1(i,l)$, $W_{t,-}^1=\sum_{i,l:Y_i[l]\neq h_t^1(x_i,l)}D_t^1(i,l)$, $D=\sum_{i=m+1}^{n}\sum_{l=1}^{L}(D_{T+1}^1(i,l))$. In Eq. (20) , only $W_{t,-}^1$ is related to $Y_i[l]$. $W_{t,-}^1$ is the weighted error rate for the tth weak hypothesis training, i.e. $W_{t,-}^1=P(h_t^1(x_i,l)\neq Y_i[l])$. If the upper error bound of co-EM comes true in Eq. (3), $P(h_t^1(x_i,l)\neq Y_i[l])$ is upper bounded by $P(h_t^1(x_i,l)\neq h_t^2(x_i,l))$. Since Eq. (20) is an increasing function of $W_{t,-}^1$, we can use $P(h_t^1(x_i,l)\neq h_t^2(x_i,l))$ to replace $P(h_t^1(x_i,l)\neq Y_i[l])$ to get the upper bound of Eq. (20). This replacement is equivalent to replace $Y_i[l]$ with sign(h_t^2 (x_i,l)). Through similar transformation, the upper bound of the fourth term on the right side of Eq. (20) can be also obtained. They are represented as follows:

$$\begin{aligned}&\sum_{i=m+1}^{n}\sum_{l}\exp(-Y_i[l]H^j(x_i,l))\\ &\quad\leq\sum_{i=m+1}^{n}\sum_{l=1}^{L}\exp(-\text{sign}(h_t^{3-j}(x_i,l))\sum_{t=1}^{T}\alpha_t^j h_t^j(x_i,l)). \qquad (21)\end{aligned}$$

Combining Eqs. (19) and (21), Theorem 1 is proved. Therefore, to minimize the training error in semi-supervised learning, we can set the labels of all samples as in Eq. (11) and minimize Z_t^1 and Z_t^2 on each round of boosting. We use this idea both in the choice of α_t and as a general criterion for the choice of weak hypothesis h_t presented in Section 3. Proof is completed. □

References

[1] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: CVPR, 2008.

[2] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos "in the wild", in: CVPR, 2009.

[3] G. Zhu, M. Yang, K. Yu, W. Xu, Y. Gong, Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor, in: ACM Multimedia, 2009.

[4] T. Xiang, S. Gong, Beyond tracking: modelling activity and understanding behaviour, International Journal of Computer Vision 67 (1) (2006) 21–51.

[5] J.C. Niebles, F.-F. Li, A hierarchical model of shape and appearance for human action classification, in: CVPR, 2007.

[6] P. Natarajan, R. Nevatia, View and scale invariant action recognition using multiview shape-flow models, in: CVPR, 2008.

[7] X. Zhu, Semi-supervised learning literature survey, in: Computer Sciences Technical Report 1530, University of Wisconsin-Madison, 2008.

[8] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Annual Workshop on Computational Learning Theory, 1998, pp. 92–100.

[9] R. Ghani, Combining labeled and unlabeled data for multiclass text categorization, in: Proceedings of the International Conference on Machine Learning, 2002.

[10] K. Nigam, R. Ghani, Analyzing the effectiveness and applicability of cotraining, in: Proceedings of the Workshop on Information and Knowledge Management, 2000.

[11] A. Levin, P. Viola, Y. Freund, Unsupervised improvement of visual detectors using co-training, in: ICCV, 2007.

[12] M. Collins, Y. Singer, Unsupervised models for named entity classification, in: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999.

[13] R. Liu, J. Cheng, H. Lu, A robust boosting tracker with minimum error bound in a co-training framework, in: ICCV, 2009.

[14] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55 (1) (1997) 19–139.

[15] R. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions, Machine Learning 37 (1999) 297–336.

[16] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: ICCV, 2005.

[17] A. Yilmaz, M. Shah, Actions sketch: a novel action representation, in: CVPR, 2005.

[18] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: PETS, 2005.

[19] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, in: ICPR, 2004.

[20] J. Sun, X. Wu, S. Yan, L.F. Cheong, T.-S. Chua, J. Li, Hierarchical spatio-temporal context modeling for action recognition, in: CVPR, 2009.

[21] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, in: ICCV, 2009.

[22] C. Fanti, L. Zelnik-Manor, P. Perona, Hybrid models for human motion recognition, in: CVPR, 2005.

[23] Y. Wang, G. Mori, Human action recognition by semi-latent topic models, Pattern Analysis and Machine Intelligence 31 (10) (2009) 1762–1774.

[24] J.C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, International Journal of Computer Vision (2008) 299–318.

[25] Y. Wang, H. Jiang, M.S. Drew, Z.-N. Li, G. Mori, Unsupervised discovery of action classes, in: CVPR, 2006.

[26] S. Yang, L. Goncalves, P. Perona, Unsupervised learning of human motion, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (7) (2003) 814–827.

[27] D. Guan, W. Yuan, Y.-K. Lee, A. Gavrilov, S. Lee, Activity recognition based on semi-supervised learning, in: International Conference on Embedded and Real-Time Computing Systems and Applications, 2007, pp. 469–475.

[28] S. Dasgupta, M. Littman, D. McAllester, Pac generalization bounds for cotraining, in: NIPS, 2001.

[29] S. Bickel, T. Scheffer, Estimation of mixture models using co-em, in: ECML, 2005.

[30] C.M. Bishop, M.E. Tipping, A hierarchical latent variable model for data visualization, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3) (1998) 281–293.

[31] H. Tang, T.S. Huang, Boosting gaussian mixture models via discriminant analysis, in: ICPR, 2008.

Tianzhu Zhang received the B.S. degree from Beijing Institute of Technology, China, in 2006. He is currently pursuing the Ph.D. degree at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. In 2009, he was an intern student in Institute for Infocomm Research, Singapore. Currently he is an intern student in China-Singapore Institute of Digital Media. His research interests include multimedia, video analysis and pattern recognition.

Si Liu received the B.E. degree from Beijing Institute of Technology, China, in 2008. He is currently pursuing the Ph.D. degree at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. In 2009, he was an intern student in Institute for Infocomm Research, Singapore. Currently he is an intern student in China-Singapore Institute of Digital Media. His research interests include multimedia, video analysis and pattern recognition.

Changsheng Xu (M'97–SM'99) received the Ph.D. degree from Tsinghua University, Beijing, China in 1996. Currently he is Professor of Institute of Automation, Chinese Academy of Sciences and Executive Director of China-Singapore Institute of Digital Media. He was with Institute for Infocomm Research, Singapore from 1998 to 2008. He was with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences from 1996 to 1998. His research interests include multimedia content analysis, indexing and retrieval, digital watermarking, computer vision and pattern recognition. He published over 150 papers in those areas. Dr. Xu is Associate Editor of ACM/Springer Multimedia Systems Journal. He served as Short Paper Co-Chair of ACM Multimedia 2008, General Co-Chair of 2008 Pacific-Rim Conference on Multimedia (PCM2008) and 2007 Asia-Pacific Workshop on Visual Information Processing (VIP2007), Program Co-Chair of VIP2006, Industry Track Chair and Area Chair of 2007 International Conference on Multimedia Modeling (MMM2007). He also served as Technical Program Committee Member of major international multimedia conferences, including ACM Multimedia Conference, International Conference on Multimedia & Expo, Pacific-Rim Conference on Multimedia, and International Conference on Multimedia Modeling.

Hanqing Lu (M‘05–SM‘06) received the Ph.D. degree in Huazhong University of Sciences and Technology, Wuhan, China in 1992. Currently he is Professor of Institute of Automation, Chinese Academy of Sciences. His research interests include image similarity measure, video analysis, object recognition and tracking. He published more than 100 papers in those areas.