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:

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

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

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

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:

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

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