Unsupervised Human Action Categorization with Consensus Information Bottleneck Method *

Xiaoqiang Yan, Yangdong Ye, Xueying Qiu
School of Information Engineering, Zhengzhou University, China
iexqyan@gmail.com, yeyd@zzu.edu.cn, iexyqiu@gmail.com

Abstract

Recent researches have shown consensus clustering can enhance the accuracy of human action categorization models by combining multiple clusterings, which can be obtained from various types of local descriptors, such as HOG, HOF and MBH. However, consensus clustering yields final clustering without access to the underlying feature representations of the human action data, which always makes the final partition limited to the quality of existing basic clusterings. To solve this problem, we present a novel and effective Consensus Information Bottleneck (CIB) method for unsupervised human action categorization. CIB is capable of learning action categories from feature variable and auxiliary clusterings simultaneously. Specifically, by performing Maximization of Mutual Information (MMI), CIB maximally preserves the information between feature variable and existing auxiliary clusterings. Moreover, to solve MMI optimization, a sequential solution is proposed to update data partition. Extensive experiments on five realistic human action data sets show that CIB can consistently and significantly beat other state-of-the-art consensus and multi-view clustering methods.

1 Introduction

Recognizing human actions automatically has been an active research area due to its wide applications, such as sports analysis, activity monitoring, action/event retrieval, etc. However, because of the large amount of intra-class variations, cluttered background and occlusion, changes of view point and camera motions, discovering action categories automatically remains a difficult and challenging task in the domain of computer vision.

In the past several years, many existing feature representations have achieved impressive progress, such as histogram of oriented gradient (HOG), histogram of optical flow (HOF) and motion boundary histograms (MBH), but their performances are still limited by the biases rooted in their self-structures. Therefore, it would be pertinent to mention feature fusion method, which has recently shown excellent performance for human action recognition. In feature fusion scheme, multiple features are integrated to produce a more discriminative new feature. For instance, [Elfiky *et al.*, 2012] investigated the optimal fusion of multiple features in the context of compact pyramid model. P. Natarajan [Natarajan *et al.*, 2012] analyzed and combined color, motion, audio and audio-visual features by late score fusion method. However, feature fusion conducts prior to the procedure of action or object recognition, which can hardly consider the relationships between classes and different features.

Recently, general frameworks to deal with multiple features have been explored. Two most common approaches are multi-view clustering and consensus clustering, and both of them assume that there is a single, true clustering of the data set. Multi-view clustering [Cao *et al.*, 2015; Kumar and Daumé, 2011; Kumar *et al.*, 2011; Wang *et al.*, 2014] uses multiple views of the data set, rather than just one view, to improve clustering performance. It directly takes multiple features as input views, and tries to fully leverage the correlative information across features. However, human actions in images and videos are usually represented by several high dimensional multi-view features, which results in the problem of the dimensionality curse. Moreover, the trade-off determination of multiple features is also a difficult problem. Consensus clustering [Iam-On *et al.*, 2011; Li *et al.*, 2007; Strehl and Ghosh, 2003; Wu *et al.*, 2015; Zhou *et al.*, 2015] aims to find a single partition of data from multiple existing basic clusterings, where each basic clustering can be generated by one type of features, such as HOG, HOF, MBH, etc. So the final clustering is a consensus from these basic clusterings. However, consensus clustering yields final clustering without access to the underlying features of the human action data, which always makes the final partition limited to the quality of existing basic clusterings.

Due to the aforementioned problems, we present a novel and effective Consensus Information Bottleneck (CIB) method. This algorithm learns action categories according to one feature variable, while taking multiple clusterings obtained from other features as auxiliary variables. The complementary information between feature variable and existing

*This work supported by the National Natural Science Foundation of China under grant No. 61170223, No. 61502434 and No. 61502432. Corresponding author: Yangdong Ye.

auxiliary clusterings is measured by Maximization of Mutual Information (MMI). To solve MMI optimization, a sequential solution is proposed to update data partition. Our experiments demonstrate the effectiveness of CIB on five realistic human action data sets. In this study, the contributions can be summarized as below:

- A novel and effective consensus information bottleneck method is proposed, which learns action categories according to one feature variable and multiple clusterings simultaneously. Thus, the proposed method can solve the overreliance of consensus clustering on existing partitions.
- An effective measurement based on MMI is designed to quantify the complementary information between feature variable and existing auxiliary clusterings.

2 Related Work

2.1 Consensus and Multi-view Clustering

In the past decades, many consensus clustering and multi-view clustering methods have been proposed. [Strehl and Ghosh, 2003] formalized consensus clustering as a combinatorial optimization problem in terms of shared mutual information. [Li *et al.*, 2007] applied non-negative matrix factorization (NMF) to clustering ensemble. [Wang *et al.*, 2011] applied a Bayesian method to consensus clustering. [Iam-On *et al.*, 2011] proposed a link-based approach to the cluster ensemble problem. [Zhou *et al.*, 2015] learned a robust consensus matrix for consensus clustering via Kullback-Leibler divergence minimization. [Wu *et al.*, 2015] provided a systematic study of K-means-based consensus clustering (KCC). Recently, several studies have shown consensus clustering can improve the performance of human action clustering. For instance, [Yang *et al.*, 2013] discovered motion primitives by hierarchical clustering optical flow in spatial and motion space. [Jones and Shao, 2014] estimated the mutual information between two clusterings and used it to improve the results of both clusterings simultaneously in the task of clustering of human action in context.

Recent advances in multi-view clustering should also be considered. [Kumar and Daumé, 2011] presented co-training and co-regularized multi-view spectral clustering. [Wang *et al.*, 2014] proposed a formulation for multi-feature clustering using minimax optimization, which unifies different feature modalities by minimizing their pairwise disagreements. [Cao *et al.*, 2015] utilized Hilbert Schmidt independence criterion as diversity term to explore the complementarity of multiple features. However, the problem of dimensionality curse and trade-off determination for multiple views needs to overcome.

2.2 Information Bottleneck

The Information Bottleneck (IB) method, introduced in [Tishby *et al.*, 1999], is an information-theoretic framework. Given the joint distribution of a source variable X and another relevant variable Y, IB tries to extract a compressed representation T of X, while preserving information about Y. The notion of compression is quantified by $I(X;T)$, while the informativeness is quantified by $I(Y;T)$. The basic quantity utilized in the IB framework is Shannon's mutual information, which is formally defined as:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \tag{1}$$

Formally, the IB function is suggested in [Tishby *et al.*, 1999] as follows:

$$\mathcal{L}_{IB}[p(t|x)] = I(X;T) - \beta I(Y;T), \tag{2}$$

where the tradeoff parameter β is the positive Lagrange multiplier controlling compression and informativeness. The solutions of this optimization problem are characterized by the bottleneck equations,

$$\begin{cases} p(t|x) = \dfrac{p(t)}{Z(x,\beta)} e^{-\beta D_{KL}(p(y|x)||p(y|t))} \\ p(y|t) = \dfrac{1}{p(t)}\displaystyle\sum_{x} p(x,y)p(t|x) \\ p(t) = \displaystyle\sum_{x,y} p(x,y,t) = \sum_{x} p(x)p(t|x) \end{cases} \tag{3}$$

where $D_{KL}(\cdot||\cdot)$ is the $Kullback{-}Leibler$ divergence, $Z(x,\beta)$ is a normalization function. The object of IB method is to find a compressed $p(t|x)$ of X.

IB has been extended to deal with multiple variables. [Slonim *et al.*, 2006] proposed multivariate extensions of IB method. [Gao *et al.*, 2007] concentrated on multi-view problem using traditional clustering ensemble, which merely generates final result from multiple basic clusterings. [Lou *et al.*, 2013] and [Yan *et al.*, 2015] proposed multi-feature extension of IB, which directly takes multiple features as input. [Xu *et al.*, 2014] utilized IB to learn a shared subspace represented by multi-view features. In this study, we estimate the complementary information between feature variable and clusterings and use it to improve the performance of final clustering.

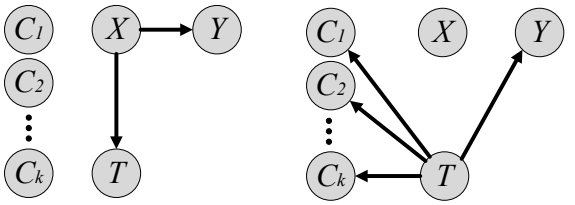
3 Consensus Information Bottleneck method

In this section, we first define the problem of consensus information bottleneck (CIB) and give one formulation of the objective function, then the optimization of CIB method is presented.

3.1 Problem Definition

Given a collection of unlabeled videos including various human actions, such as cycling, boxing, diving, etc. In realistic scenarios, the performances of single feature representation are still limited by the biases rooted in their self-structures. Two most common approaches dealing with multiple features are consensus clustering and multi-view clustering, but they still have their own flaws, for instance, the overreliance of consensus clustering method on existing partitions, the dimensionality curse and trade-off determination of multi-view clustering. So we are curious about whether we can draw the advantage from both consensus and multi-view clustering.

Suppose there is an unlabeled video collection X taking values from $\{x_1, x_2, \ldots, x_m\}$, which contains various action



(a) Data Compression (b) Information preservation

Figure 1: The model of CIB method. (a) The data compression shows the compressing relationships among multiple variables. X, Y and T denotes the source, relevant and compression variable respectively; $C_1, \ldots, C_k$ are multiple valid clusterings generated from multiple features. (b) The information preservation implies that the compressed variable T should preserve the information with respect to feature variable Y and auxiliary clusterings $C_1, \ldots, C_k$.

categories, and m is the total number of videos in the data. We assume that there are $k+1$ discrete random variables $Y_1, \ldots, Y_{k+1}$ on behalf of $k+1$ types of features of the video data, such as HOG, HOF, MBH, STIP, 3DSIFT, etc. First, we stochastically choose one discrete random variable as feature variable $Y = \{y_1, y_2, \ldots, y_n\}$, which characterizes the videos X from one cue. Then the remaining k variables are used to generate multiple valid clusterings $C_1, \ldots, C_k$ of the videos. The task of CIB algorithm is to learn a compressed representation $p(t|x)$ from both feature variable and multiple clusterings.

3.2 Objective Function

In this section, a novel and effective method named CIB is proposed to discover action categories in videos by considering both feature variable and multiple clusterings. The model of CIB is given in Figure 1, which has two parts: data compression and information preservation. In data compression, original videos X including various actions are compressed into variable T. In information preservation part, the compressed variable T should preserve the information about feature variable Y and existing clusterings $C_1, \ldots, C_k$. The objective function of CIB can be formulated as follows according to the Equation 2:

$$\begin{aligned}&\mathcal{L}_{min}[p(t|x)] = I(X;T) - \\ &\beta \cdot [I(Y;T) + \lambda_1 I(C_1;T) + \cdots + \lambda_k I(C_k;T)],\end{aligned} \tag{4}$$

where $I(X;T)$ measures the compactness of source variable X into the compressed variable T. $I(Y;T) + I(C_1;T), \ldots, I(C_k;T)$ measures what information the compressed variable T should preserve, of which the $I(Y;T)$ is on behalf of its relevant feature information, $I(C_1;T), \ldots, I(C_k;T)$ denotes its information with respect to the information of existing clusterings. β is the balance parameter controlling the trade-off between information compression and preservation. $\lambda_1, \ldots, \lambda_k$ are trade-off parameters to balance the influence of different existing clusterings obtained from multiple features. For the convenience of optimization, we consider the problem of maximizing

$$\begin{aligned}&\mathcal{L}_{max}[p(t|x)] = [I(Y;T) + \lambda_1 I(C_1;T) + \cdots + \\ &\lambda_k I(C_k;T)] - \beta^{-1} \cdot I(X;T),\end{aligned} \tag{5}$$

which is clearly equivalent to minimize the Equation 4 by dividing $-\beta$. In the task of unsupervised human action categorization, the number of action categories M is much less than the source video variable X, which implies a significant compression. Therefore, we concentrate on preserving the information of T with respect to feature variable and existing clusterings maximally. To achieve this goal, the value of β is fixed to ∞, so the mutual information $I(X;T)$ is eliminated. Now, the objective function of CIB can be rewritten as

$$\mathcal{L}_{max}[p(t|x)] = I(Y;T) + \lambda_1 I(C_1;T) + \cdots + \lambda_k I(C_k;T) \tag{6}$$

the fraction $I(Y;T)$ indicates to maximally preserve relevant feature information they capture about Y, $\lambda_1 I(C_1;T), \ldots, \lambda_k I(C_k;T)$ denote the preserved information with respect to existing clusterings. This study focuses on the hard clustering, where the value of $p(t|x)$ is either 0 or 1. Now, the task of our unsupervised human action categorization is reduce to maximize the objective function 6.

3.3 Optimization

To find an optimal partition of source videos X, a sequential draw-and-merge optimization method is presented. The sequential draw-and-merge method starts with random partition of X into M clusters. At each step, a potential $x \in X$ is drawn from its current cluster t^{old}, and then represented as a new singleton cluster $\{x\}$. Now, we have $M+1$ clusters. To ensure the total number of clusters is M, we must merge the singleton cluster $\{x\}$ into one of clusters t^{new}. So we should guarantee to increase the value of objective function 6 at each draw-and-merge procedure.

The key issue of CIB is to decide which cluster t^{new} the singleton cluster $\{x\}$ should be merged into. Let $\mathcal{L}^{bef}$ and $\mathcal{L}^{aft}$ denote the value of function 6 before and after x is drawn from its current cluster; let $\mathcal{L}^{new}$ denote the value of function 6 after x is merged into some new cluster t^{new}. The measure of deciding merger procedure is called "merger cost" $d_{\mathcal{L}}$, which is on behalf of the value change after one draw and merge procedure, i.e. $d_{\mathcal{L}} = \mathcal{L}^{aft} - \mathcal{L}^{new}$. We should merge $\{x\}$ into t^{new} such that $t^{new} = \arg\min d_{\mathcal{L}}$. In the following, we will give the solution to this problem.

Now, we first calculate the difference of $I(Y;T)$ in function 6 between the values of $\mathcal{L}^{new}$ and $\mathcal{L}^{aft}$, which denoted by $\Delta I_{feature}$. Let x be merged into cluster t and become a new cluster $\widetilde{t}$, i.e. $\{\{x\}, t\} \Rightarrow \widetilde{t}$, then

$$p(\widetilde{t}) = p(x) + p(t), \tag{7}$$

$$p(y|\widetilde{t}) = \frac{p(x)}{p(\widetilde{t})} p(y|x) + \frac{p(t)}{p(\widetilde{t})} p(y|t). \tag{8}$$

So,

$$\Delta I_{feature} = I(T^{aft};Y) - I(T^{new};Y)$$
$$= p(x)\sum_y p(y|x)\frac{p(y|x)}{p(y)} + p(t)\sum_y p(y|t)\frac{p(y|t)}{p(y)}$$
$$-p(\widetilde{t})\sum_y p(y|\widetilde{t})\frac{p(y|\widetilde{t})}{p(y)}.$$

Using Equation 7 and Equation 8, we can get the following results.

$$\Delta I_{feature} = p(x)\sum_y p(y|x)\frac{p(y|x)}{p(y)} + p(t)\sum_y p(y|t)\frac{p(y|t)}{p(y)}$$
$$-\sum_y p(x)p(y|x)\frac{p(y|\widetilde{t})}{p(y)} - \sum_y p(t)p(y|t)\frac{p(y|\widetilde{t})}{p(y)}$$
$$= p(x)\sum_y p(y|x)\frac{p(y|\widetilde{t})}{p(y)} + p(t)\sum_y p(y|t)\frac{p(y|t)}{p(y|\widetilde{t})}$$
$$= p(x)D_{KL}[p(y|x)||p(y|\widetilde{t})] + p(t)D_{KL}[p(y|t)||p(y|\widetilde{t})]$$
$$= [p(x)+p(t)]JS_{\prod}[p(y|x),p(y|t)],$$

where the $JS_{\prod}$ is the $Jensen$-$Shannon$ divergence, $\prod = \{\frac{p(x)}{p(x)+p(t)}, \frac{p(t)}{p(x)+p(t)}\}$. Since $JS_{\prod} \geq 0$, we obtain $\Delta I_{feature} \geq 0$.

Next, we give the calculation of $\lambda_1 I(C_1;T) + \cdots + \lambda_k I(C_k;T)$. Suppose C_i is the i'th clustering from one feature representation, so we can compute the confusion matrix $R(p,q)$ of C_i and T, where p and q are clusters in C_i and T. The elements in R are co-occurrence number of source data x belongs to both p in C_i and q in T. So the mutual information between T and C_i can be calculated. Now, the difference of $\lambda_1 I(C_1;T) + \cdots + \lambda_k I(C_k;T)$ in function 6 between the values of $\mathcal{L}^{aft}$ and $\mathcal{L}^{new}$ is calculated by $\Delta I_{clustering} = \lambda_1[I(C_1;T^{aft}) - I(C_1;T^{new})] + \cdots + \lambda_k[I(C_k;T^{aft}) - I(C_k;T^{new})]$. So the total merger cost $d_{\mathcal{L}}$ is calculated as follows:

$$d_{\mathcal{L}} = \Delta I_{feature} + \Delta I_{clustering}. \quad (9)$$

At each draw-and-merge step, x will be merged into t^{new} such that $t^{new} = \arg\min d_{\mathcal{L}}$, where $d_{\mathcal{L}}$ is on behalf of the information loss. Note that, once x is merged into one new cluster, there must be some information loss, so we can get $\Delta I_{clustering} \geq 0$ and $d_{\mathcal{L}} \geq 0$. Assumed X has a true clustering C, we can obtain $I(T;C_i) \leq I(C;C_i)$. And because T is a compressed representation of X, so $I(T;Y) \leq I(X;Y)$. So the value of objective function 6 is upper bounded, which is guaranteed to converge to a stable solution. The details of CIB algorithm is shown in Algorithm 1.

3.4 Complexity Analysis

In this section, we give the computation costs of the CIB algorithm. The time complexity of the random initialization at step 2 and the procedure of drawing some x at step 5 is $O(N)$. At step 6, the merger cost $d_{\mathcal{L}}$ between x and each new cluster t should be calculated, which takes $O(lM|X||Y|)$, where l is the number of repetitions that should be performed over X until convergence is attained. From the experimental section, the CIB takes a few repetitions to converge. The M is the number of clusters, usually, it can be considered as constant. Note that the calculation of mutual information between t^{new} and multiple clusterings $C_1,\ldots,C_k$ takes $O(1)$, so the time complexity of CIB is $O(|X||Y|)$.

Algorithm 1 The Consensus Information Bottleneck: CIB

1: **Input:** Joint distribution $p(X,Y)$; multiple clusterings $C_1,\ldots,C_k$; trade-off parameters $\lambda_1,\cdots,\lambda_k$, cluster number M).
2: **Initialize:** $T \leftarrow$ Random partition of X into M clusters;
3: **repeat**
4: **for** For every $x \in X$ **do**
5: **Draw:** Remove x from current cluster $t(x)$;
6: **Computing merger cost:**
For data point x, calculate merger costs $d_{\mathcal{L}}$ of all possible reassignments of x to different clusters based on Equation 6;
7: **Merge:** Merge x into cluster t^{new} such that $t^{new} = \arg\min_{t\in\mathcal{T}} d_{\mathcal{L}}$;
8: **end for**
9: **until** Convergence
10: **Output:** A partition T of X into M clusters.

4 Experiments

4.1 Data Set Descriptions

In this section, to evaluate the effectiveness of the proposed CIB algorithm, extensive experiments are conducted on five benchmark video data sets. The Weizmann data set contains 10 action categories performed by 9 people, to provide a total of 90 videos. The KTH data set contains 6 types of human actions performed by 25 subjects in outdoor and indoor environment, total of 599 sequences. UCF Sports [Rodriguez *et al.*, 2008] data set consists of 1100 videos of various sports action videos, taken from various broadcast sources. UCF50 [Reddy and Shah, 2013] is an action recognition data set with 50 action categories, consisting of 6000 realistic videos in YouTube. HMDB [Kuehne *et al.*, 2011] data set is a recently released large video database, which has been collected from various sources, mostly movies, and contains 6849 clips divided into 51 action categories. Figure 2 shows example frames in HMDB data set.

Figure 2: Video frames of example action classes in HMDB data set.

4.2 Experimental Setup

To incorporate the consensus information of auxiliary clusterings, we adopt the following three descriptors: STIP,

Data sets	IB			IB	IB	k-means	pLSA	LDA	NCuts	CIB
	HOG	HOF	STIP	Best	Con					
Weizmann	59.5	69.8	66.5	69.8	67.9	52.8	61.2	64.2	64.5	**76.4** (↑)
KTH	58.5	68.5	68.9	68.9	70.3	52.0	66.4	66.9	68.2	**72.1** (↑)
UCF Sports	38.7	53.8	50.4	53.8	52.4	40.3	46.3	51.7	47.1	**55.8** (↑)
UCF50	33.1	34.0	31.2	34.0	34.2	29.0	30.3	29.6	31.7	**39.2** (↑)
HMDB	19.0	22.3	21.8	22.3	22.8	13.7	18.9	20.6	19.0	**23.3** (↑)
Average	41.8	49.7	47.8	49.8	49.5	37.6	44.6	46.6	46.1	53.3

Table 1: AC (%) comparisons of CIB with original IB and other four traditional clustering methods.

HOG, HOF, to extract motion representations of the actions. Then the bag-of-words (BoW) model is adopted to represent videos. The size of the vocabulary in BoW model is set to 1000, which results in a 1000 dimensional frequency histogram of motion features. We choose one feature from STIP, HOG, HOF randomly as our feature variable, the remaining two features are naturally treated as clustering variables to construct auxiliary clusterings. In this paper, the clustering accuracy (AC) [Cai *et al.*, 2009] is employed to evaluate the performance of different methods. The number of categories M is set to be identical with number of real categories on each data set. As all algorithms are stochastic, all experiments are run 10 times, and we report the average clustering results.

4.3 Experimental Results and Analysis

To validate the performance of the CIB approach, we adopt five types of baselines: original IB method, traditional clustering methods, consensus clustering methods, multi-view clustering methods action clustering methods.

Original IB and CIB

The original IB method can only process one feature variable. As an extension of IB method, CIB can handle feature variable and auxiliary clusterings simultaneously in the clustering procedure. In this section, we conduct the experiments to compare the performance of CIB and original IB. As illustrated in Table 1, we can get the following observations:

- The performances of IB method on HOG, HOF, STIP are different. For instance, IB method obtains the best AC in terms of HOF feature (69.8%) on Weizmann data set, while it gets the best result in STIP feature (68.9%) on KTH data set. This is mainly because no feature can perform consistently well for different clustering tasks.
- The performances of original IB method on concatenated feature are less than satisfactory. For instance, there is a decline on Weizmann and UCF Sports (1.9% and 1.4%). So, concatenating features simply can not consistently attain improved results compared with individual feature.
- The benefits of CIB method are verified as shown in Table 1. The CIB algorithm obtains improvement on all data sets used in this study in terms of AC value (6.6%, 3.2%, 2.0%, 5.2%, 1.0% respectively) compared with the best results of original IB on three individual features. So we can get the conclusion that the CIB algorithm can consistently improve the clustering performance compared with the best results of original IB on individual feature.

As for the comparison with pLSA, LDA, NCuts, the results of CIB method obtain great improvement (15.7%, 8.7%, 6.7%, 7.2% respectively). This phenomenon demonstrates that the proposed CIB method is effective by exploiting the complementary impact of feature and multiple clusterings.

Consensus Clustering Methods and CIB

CIB method gives a better performance increase over the baseline consensus clustering methods, such as CSPA [Strehl and Ghosh, 2003], MCLA [Strehl and Ghosh, 2003], BCE [Wang *et al.*, 2011]. We run IB method 15 times with different initializations to obtain 15 base clusterings, all the consensus clustering methods are then applied on them. From Table 2, we note that the performance improvements of CIB over three consensus methods on five data sets are 2.4%, 2.0%, 4.9% in terms of AC, respectively. This is mainly because that CIB can solve the overreliance of consensus clustering on existing clusterings.

Multi-view Clustering Methods and CIB

In this section, the experiments are conducted to verify the effectiveness of CIB method compared with multi-view clustering methods: the Co-Training multi-view Spectral Clustering(CTSC) [Kumar *et al.*, 2011], the Co-Regularized multi-view Spectral Clustering (CRSC) [Kumar and Daumé, 2011] and the Robust Multi-view Spectral Clustering (RMSC) [Xia *et al.*, 2014]. We implement the methods following the original works. A glance at Table 2 shows that our method is always better than the other three multi-view clustering methods by a large margin. Note that, CIB method just adopts one type of feature as input, which will reduce multi-view data dimension significantly.

Action Clustering Methods and CIB

For the comparison with action clustering in videos, we adopt DAKM [Jones and Shao, 2014], MfIB [Lou *et al.*, 2013] and MvIB [Yan *et al.*, 2015] as baselines. DAKM estimates the mutual information between clusterings and used it to improve the results of each clustering simultaneously. MfIB and MvIB are extensions of original IB and multivariate IB respectively, both of them take multiple features as input directly and try to leverage the correlative information across features. Differently, (1) CIB learns action categories according to one feature variable and multiple clusterings other than multiple features; (2) CIB is a consensus extension of original IB, which can capture the essence of consensus clustering

Data sets	Consensus clustering			Multi-view clustering			Action clustering			CIB
	CSPA	MCLA	BCE	CTSC	CRSC	RMSC	DAKM	MfIB	MvIB	
Weizmann	76.3	75.3	68.3	48.6	47.3	43.4	66.3	67.4	68.9	76.4
KTH	68.4	69.8	64.7	63.7	62.1	71.2	70.1	70.3	71.9	72.1
UCF Sports	54.6	59.8	54.4	53.2	54.6	47.7	53.9	54.1	54.6	55.8
UCF50	31.7	32.2	32.0	34.1	31.5	33.4	34.5	33.7	34.1	39.2
HMDB	23.7	19.2	22.7	22.6	22.9	22.7	20.1	21.3	22.1	23.3
Average	50.9	51.3	48.4	44.4	43.7	43.7	49.0	50.0	50.3	53.3

Table 2: AC (%) comparisons of CIB with other three types of clustering methods.

and multi-view clustering. From Table 2, the performances of CIB are better than the action clustering methods.

4.4 The Impact of Parameters

Since there are two auxiliary clusterings (C_1, C_2) in our experiments, we show the impacts of varying parameter λ_1 and λ_2 on the performance of CIB on different data sets in Figure 3. Set $\lambda_1 + \lambda_2 = 1$, where λ_1 acts on C_1, λ_2 acts on C_2. Then, vary the values of λ_1 from 0 to 1, with 0.1 as the gap between adjacent values. When $\lambda_1 = 0$, the CIB method only acts on feature variable Y and clustering C_2. As we can see from Figure 3, we get the following observations:

- The performance of CIB fluctuates to some extent according to the trade-off parameters. When λ_1 varies from 0 to 1, the value of λ_2 varies from 1 to 0 correspondingly. However, the AC values of CIB change in small range on all the five data sets in Figure 3. That is to say the proposed algorithm is not sensitive to the trade-off parameters.
- Except two values on Weizmann data set, the performance of CIB is always better than the best results of original IB on individual feature. This phenomenon demonstrates that the CIB method can cope with feature variable and auxiliary clusterings effectively and it is relatively easy to choose the trade-off parameters.

4.5 Convergence of CIB algorithm

Figure 4 shows the repetitions of CIB on UCF50 and HMDB data set. We observe that the values of objective function 6 increase monotonically with each repetition and 14 iterations are enough for convergence.

5 Conclusions

In this paper, a novel and effective Consensus Information Bottleneck (CIB) method has been introduced for learning action categories from feature variable and auxiliary clusterings simultaneously. CIB adopts maximization of mutual information to measure feature and clusterings, which can solve the overreliance of consensus clustering on existing base clusterings. Therefore, the compressed results can reflect the hidden patterns by multiple cues. The experiments on five benchmark human action data sets have confirmed the effectiveness of the proposed CIB algorithm. Our further work will focus on incorporating heterogeneous features into the CIB approach, and applying it to more difficult tasks, such as action and scene categorization on large-scale data.

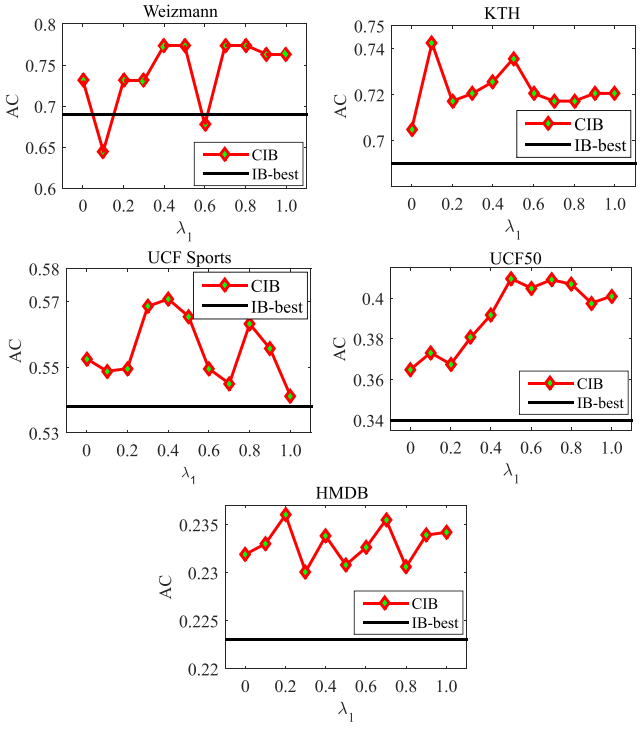


Figure 3: Performance of CIB with various λ.

References

[Cai *et al.*, 2009] Deng Cai, Xuanhui Wang, and Xiaofei He. Probabilistic dyadic data analysis with local and global consistency. In *International Conference on Machine Learning (ICML)*, pages 105–112, 2009.

[Cao *et al.*, 2015] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–594, 2015.

[Elfiky *et al.*, 2012] Noha M Elfiky, Fahad Shahbaz Khan, and Joost Van De Weijer. Discriminative compact pyramids for object and scene recognition. *Pattern Recognition (PR)*, 45(4):1627–1636, 2012.

[Gao *et al.*, 2007] Yan Gao, Shiwen Gu, Jianhua Li, and Zhining Liao. The multi-view information bottleneck clus-

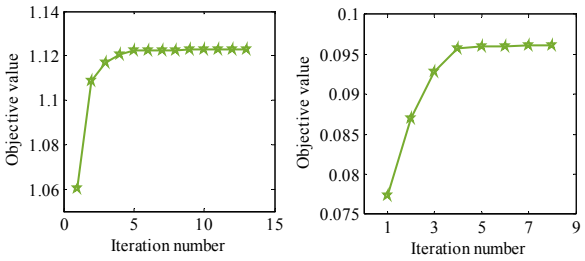


Figure 4: The iterations of CIB on UCF50 (left) and HMDB (right) data set.

tering. In *Advances in Databases: Concepts, Systems and Applications*, pages 912–917. 2007.

[Iam-On *et al.*, 2011] Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price. A link-based approach to the cluster ensemble problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(12):2396–2409, 2011.

[Jones and Shao, 2014] Simon Jones and Ling Shao. Unsupervised spectral dual assignment clustering of human actions in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 604–611, 2014.

[Kuehne *et al.*, 2011] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2556–2563, 2011.

[Kumar and Daumé, 2011] Abhishek Kumar and Hal Daumé. A co-training approach for multi-view spectral clustering. In *International Conference on Machine Learning (ICML)*, pages 393–400, 2011.

[Kumar *et al.*, 2011] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1413–1421, 2011.

[Li *et al.*, 2007] Tao Li, Chris Ding, Michael Jordan, et al. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *IEEE International Conference on Data Mining (ICDM)*, pages 577–582, 2007.

[Lou *et al.*, 2013] Zhengzheng Lou, Yangdong Ye, and Xiaoqiang Yan. The multi-feature information bottleneck with application to unsupervised image categorization. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1508–1515, 2013.

[Natarajan *et al.*, 2012] Pradeep Natarajan, Shuang Wu, and Shiv Vitaladevuni. Multimodal feature fusion for robust event detection in web videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1298–1305, 2012.

[Reddy and Shah, 2013] Kishore K Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vison and Applications (MVA)*, 24(5):971–981, 2013.

[Rodriguez *et al.*, 2008] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[Slonim *et al.*, 2006] Noam Slonim, Nir Friedman, and Naftali Tishby. Multivariate information bottleneck. *Neural Computation*, 18(8):1739–1789, 2006.

[Strehl and Ghosh, 2003] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research (JMLR)*, 3(3):583–617, 2003.

[Tishby *et al.*, 1999] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Annual Allerton Conference on Communnication, Control and Computing*, pages 368–377, 1999.

[Wang *et al.*, 2011] Hongjun Wang, Hanhuai Shan, and Arindam Banerjee. Bayesian cluster ensembles. *Statistical Analysis and Data Mining (SADM)*, 4(1):54–70, 2011.

[Wang *et al.*, 2014] Hongxing Wang, Chaoqun Weng, and Junsong Yuan. Multi-feature spectral clustering with minimax optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4106–4113, 2014.

[Wu *et al.*, 2015] Junjie Wu, Hongfu Liu, Hui Xiong, Jie Cao, and Jian Chen. K-means-based consensus clustering: A unified view. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 27(1):155–169, 2015.

[Xia *et al.*, 2014] Rongkai Xia, Yan Pan, Lei Du, and Jian Yin. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *American Association for Artificial Intelligence(AAAI)*, pages 2149–2155, 2014.

[Xu *et al.*, 2014] Chang Xu, Dacheng Tao, and Chao Xu. Large-margin multi-view information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(8):1559–1572, 2014.

[Yan *et al.*, 2015] Xiaoqiang Yan, Yangdong Ye, and Zhengzheng Lou. Unsupervised video categorization based on multivariate information bottleneck method. *Knowledge-Based Systems (KBS)*, 84:34–45, 2015.

[Yang *et al.*, 2013] Yang Yang, Saleemi Imran, and Shah Mubarak. Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. *IEEE Transactions on Pattern Analysis Machine Intelligence (TPAMI)*, 35(7):1635–1648, 2013.

[Zhou *et al.*, 2015] Peng Zhou, Liang Du, Hanmo Wang, Lei Shi, and Yi-Dong Shen. Learning a robust consensus matrix for clustering ensemble via kullback-leibler divergence minimization. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4112–4118, 2015.