Structured Feature Selection

Tian Gao Ziheng Wang Qiang Ji
Department of ECSE, Rensselaer Polytechnic Institute, USA
{gaot, wangz10, jiq}@rpi.edu

Abstract

Feature dimensionality reduction has been widely used in various computer vision tasks. We explore feature selection as the dimensionality reduction technique and propose to use a structured approach, based on the Markov Blanket (MB), to select features. We first introduce a new MB discovery algorithm, Simultaneous Markov Blanket (STMB) discovery, that improves the efficiency of state-of-the-art algorithms. Then we theoretically justify three advantages of structured feature selection over traditional feature selection methods. Specifically, we show that the Markov Blanket is the minimum feature set that retains the maximal mutual information and also gives the lowest Bayes classification error. Then we apply structured feature selection to two applications: 1) *We introduce a new method that enables STMB to scale up and show the competitive performance of our algorithms on large-scale image classification tasks.* 2) *We propose a method for structured feature selection to handle hierarchical features and show the proposed method can lead to big performance gain in facial expression and action unit (AU) recognition tasks.*

1. Introduction

Large scale computer vision problems become increasingly relevant in the age of big data. Recognition tasks in these vision problems are often aided by large-scale features. Many such high-dimensional features, such as Fisher Vector (FV) [33], Vector of Locally Aggregated Descriptors (VLAD) [18], and Super Vectors [49], have been proposed for different visual recognition and image retrieval tasks, and they have achieved state-of-the-art performance. However, these large dimensional features significantly increase computational complexity as well as memory requirement. If one chooses to adopt schemes such as spatial pyramid matching structure [5], the feature size becomes even larger. As a result, dimensionality reduction techniques such as feature compression [35, 13] have been widely used to alleviate the computational burden. In particular, Product Quantization (PQ) methods [17, 27, 12] have achieved a lot of success in the recognition tasks [6]. Other dimension reduction techniques such as feature extraction and subspace learning [37, 26, 28, 43] are also regularly used by the vision community. Recently, studies [48] have shown that feature selection may perform equally well or better for certain vision tasks, as the linear projection assumption in feature compression methods can be easily violated in real domains. Furthermore, for well-defined features such as human body joint positions and facial component locations, projection-based methods would lose features' physical meanings.

In this work, we explore feature selection as the dimensionality reduction technique and propose a structured feature selection approach based on Markov Blanket (MB) discovery. From the graph-theoretic point of view, the Markov Blanket of a variable consists of its immediate neighbors in a graphical model. Probabilistically, the MB has an unique and valuable property: given the MB of a target node, all other nodes are independent of the target node. Given this property of the MB, feature selection for a target variable can be formulated as finding the features that are the MB of the target node. Feature selection can therefore be carried out as the MB discovery for the target variable such as class labels. Unlike the traditional feature selection methods, feature selection using the MB considers the structural information among variables. We introduce a new MB discovery algorithm that improves the efficiency over the existing MB discovery method. We also theoretically show that the optimal Markov Blanket is the minimum feature set that retains the maximal amount of mutual information to the target node, and also gives the lowest Bayes classification error. We then apply the proposed MB-based method to large-scale and hierarchical feature selection in computer vision tasks to show its effectiveness.

2. Related Work

Feature selection have been applied in many vision tasks such as face recognition [15], eye movement analysis [4], medical imaging [50], image annotation [20], object and scene recognition [34], and digit and texture classification [7]. Their methods can be roughly divided into three different approaches: the wrapper, embedded, and filter approaches. The wrapper approach uses the classification accuracy as the criterion to select features. The embedded approach uses a classifier-specific criterion to select features. In comparison, filter methods select features by using only the statistics of the data. They make the least amount of assumption, are classifier-independent, are the easiest to scale up, and can be used in conjunction with any supervised learning models. For this work, we will focus on the filter-based approach.

Although most feature selection methods try to achieve the best classification accuracy, they do not explicitly consider structure relationships among features. Structure relationships are the essence of human reasoning and decision-making. If we know the relationships among several features or factors, we can then predict the consequences of certain actions [29]. This motivates us to use MB-based *structured feature selection* methods[1]. They are a type of filter feature selection methods. Many principled algorithms have been proposed to find the Markov Blanket. The Koller and Sahami algorithm (KS) [19] is the first method to find the MB via an approximated search. Since then, many popular methods such as HITON-MB [2] and Iterative Parent-Child based search of MB (IPCMB) [10] algorithms have been proposed to improve both the accuracy and efficiency of MB discovery. Almost all of them use independence tests to infer the MBs and have been successfully applied to feature selection [19, 41, 31] in the field of drug discovery, clinical diagnosis, text categorization, document collection, and many machine learning datasets. Extensive experiments [3] show the extremely competitive performance of MB-based structured feature selection.

Although structured feature selection considers the structure information, it is very different from traditional structural pattern recognition [1, 47]. Structural or Syntactic pattern recognition considers the structure in physical shapes to construct either a grammar or isomorphic graphs. The nodes in the graph are deterministic and classification relies on graph-matching techniques.

3. Structured Feature Selection

3.1. Background

We use the standard definitions for probabilities, entropy, and mutual information. Let $\mathbf{V}$ denote a set of random variables. A Bayesian Network for $\mathbf{V}$ is represented by a pair (G, θ). The structure G is a directed acyclic graph (DAG) with nodes corresponding to the random variables in $\mathbf{V}$ and directed links capturing the dependencies between the connected nodes. If a directed link exists from node X to node Y, X is a *parent* of Y and Y is a *child* of X. Two non-adjacent nodes that share the same children are *spouses*. *Descendants* of T include T's children, children's children and so on. The parameters θ represent the conditional probability distribution of each node $X \in \mathbf{V}$ given its parents. For the rest of the paper, we use $|\cdot|$ to represent the size of a set, and $X \perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y$ to represent independence and dependence between X and Y, respectively.

In a Bayesian network (BN), the *Markov Blanket* [29] of a target variable T, $\mathbf{MB}_T$, is the minimal set of nodes conditioned on which all other nodes are independent of T, denoted as $X \perp\!\!\!\perp T|\mathbf{MB}_T, \forall X \subseteq \{\mathbf{V} \setminus T \setminus \mathbf{MB}_T\}$. $\mathbf{MB}_T$ is minimal if none of its proper subsets satisfies the above property. In a BN, the MB of a node consists of the node's parents, children, and spouses. Given its MB, all paths from the target node to the remaining nodes are cut off and all the remaining nodes become irrelevant to the target node. For example, in Figure 1, the MB of node T, $\mathbf{MB}_T$, consists of its parent node P, its child node C, and its spouse S. All other nodes A, B, R, D, and E are independent of T, given $\mathbf{MB}_T$. Given an unknown distribution $\mathcal{P}$ that satisfies the Markov condition with respect to an unknown DAG G, Markov Blanket discovery is the process used to estimate the MB of a target node from independently and identically distributed (i.i.d) data samples D of $\mathcal{P}$. Assuming the faithfulness condition holds and independence tests correctly reflect independence, the MB of a target node is uniquely identifiable.

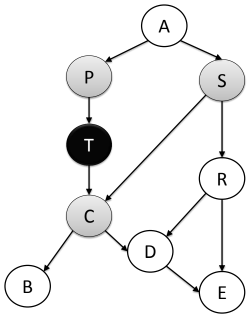

Figure 1. Sample Bayesian Network. Black node T is the target node and the shaded nodes are the Markov Blanket of T.

Lastly, one of the main concepts in MB algorithms is the symmetry constraint, which states that for a node X to be a parent or child of T, both of the following statements must hold true: X must be in the parent and children (PC) set of T and T must be in the PC set of X, i.e. $X \in \mathbf{PC}_T$ and $T \in \mathbf{PC}_X$. Recent state-of-the-art algorithms [2, 10] employ the symmetry constraint to remove those false positive PC nodes in the returned PC set. In Figure 1, for example, using the IPCMB algorithm, node D would be in the returned PC set of T due to $D \not\perp\!\!\!\perp T|\mathbf{Z}, \forall \mathbf{Z} \subseteq \{P, C\}$, and D can be removed using the symmetry constraint.

[1] Existing work sometimes also refers to MB-based feature selection as Causal Feature Selection [3].

3.2. MB Discovery Algorithms

Existing Markov Blanket discovery algorithms typically first identify the PC nodes of the target node and then the spouse nodes. The detailed description of previous works can be found in the supplementary material. We propose a new Simultaneous Markov Blanket (STMB) discovery algorithm, building on the state-of-the-art algorithm IPCMB [10] to improve MB discovery efficiency. Compared with IPCMB, STMB does not use the very costly symmetry constraint that requires to find the MBs of each PC node of the target. STMB is based on the BN and MB topology, and systemically performs independence tests to identify the PC set of a target node first and then find the spouses. The proposed algorithm, shown in the Algorithm 1, has two steps; 1) the first step of STMB identifies the PC set using the same step 1 as IPCMB. It returns the learned PC set of T and separation sets for each non-PC variable X (i.e., $X \not\perp\!\!\!\perp T|\mathbf{Sep}_T\{X\}$). 2) in Step 2, STMB finds the spouses and removes the non-MB descendants from the PC set at the same time. Previous finding concludes that potential non-MB descendants may exist in the identified PC set in Step 1 [31], and we introduce a new method instead of the symmetry constraint to remove them. Specifically in Step 2 of Algorithm 1, STMB looks for a node $Y \in \mathbf{PC}_T$ that unblocks one path from T to some node $X \in \mathbf{V} \backslash \mathbf{PC}_T$ (i.e., a candidate spouse set). If such a Y was found, X could be a spouse (Line 13) and Y could be a child or non-MB descendant node. If Y is a non-MB descendant (Line 9), Line 10 removes Y from $\mathbf{PC}_T$. Then starting at Line 19, STMB tests for false positive spouses X by conditioning on other nodes that are unblocked by each child Y. If X and T are independent, X is removed from the spouse candidate set (Line 23). STMB then tests for other non-MB descendants X in the PC set that may have multiple unblocked paths to the target. If X and T are conditionally independent, X is removed from the PC set (Line 29).

STMB reduces the worst-time complexity of the state-of-the-art algorithm IPCMB [10] by $O(C)$, where $C = \max_i |\mathbf{PC}_i|, i \in \mathbf{PC}_T$. Overall, STMB belongs to standard constraint-based BN learning procedures [30]. For more detailed analysis and empirical evaluation of STMB on standard machine learning datasets, please refer to the supplementary material. The soundness and completeness of STMB can be proven from the algorithm procedure.

Algorithm 1 STMB Algorithm

```
1:  Input: Data, $D$; target node, $T$
2:  CanMB$_T \leftarrow \mathbf{V} \setminus T$;
    {step 1: find the PC set }
3:  $[\mathbf{PC}_T, \mathbf{Sep}_T] \leftarrow$ RecogPC($T$,CanMB$_T$,$D$);
    {step 2: find spouses and remove descendants}
4:  $\mathbf{spouse}_T \leftarrow \emptyset$;
5:  $\mathbf{remove} \leftarrow \emptyset$;
6:  for each $Y \in \mathbf{PC}_T$ do
7:    for each $X \in \{$ CanMB$_T \setminus \mathbf{PC}_T\}$ do
8:      if $X \not\perp\!\!\!\perp T|\mathbf{Sep}_T\{X\} \cup Y)$ then
9:        if $Y \perp\!\!\!\perp T|\mathbf{Z}, \mathbf{Z} \subseteq \{\mathbf{PC}_T \cup X \setminus Y\}$ then
10:         $\mathbf{remove} \leftarrow \mathbf{remove} \cup Y$;
11:         break;
12:       else
13:         $\mathbf{spouse}_T\{Y\} \leftarrow \mathbf{spouse}_T\{Y\} \cup X$;
14: $\mathbf{PC}_T \leftarrow \mathbf{PC}_T \setminus \mathbf{remove}$;
15: for each $Y$ in $\mathbf{spouse}_T$ do
16:   for each $S$ in nonempty $\mathbf{spouse}_T\{Y\}$ do
17:     $\mathbf{testSet} \leftarrow \mathbf{PC}_T \cup \mathbf{spouse}_T\{Y\} \setminus S$;
18:     if $S \perp\!\!\!\perp T|\mathbf{testSet}$ then
19:       remove $S$ from $\mathbf{spouse}_T\{Y\}$;
20: for each $X \in \mathbf{PC}_T$ do
21:   if $X \perp\!\!\!\perp T|\{\mathbf{PC}_T \cup \mathbf{spouse}_T \setminus X\}$ then
22:     $\mathbf{PC}_T \leftarrow \mathbf{PC}_T \setminus X$;
23: $\mathbf{MB} \leftarrow \mathbf{spouse}_T \cup \mathbf{PC}_t$;
```

3.3. Properties of MB-based Feature Selection

MB discovery can be thought of as an information-theoretic filter approach (i.e., classifier independent) to select features, as it only tests independence relationships among features. Compared to standard filter feature selection methods, structured feature selection using MB have several advantages [3]: 1) it guarantees the selection optimality, 2) it returns the minimum feature set, 3) and it has minimum Bayes classification error. Although empirical studies [14, 3] on machine learning datasets support the superiority of MB methods, there are no current theoretical justifications on these advantages, besides Koller & Sahami [19] who showed the first advantage of the optimality by minimizing cross entropy. Aiming to enrich the theoretical analysis, we first show the feature selection optimality from the mutual information point of view, in preparation to justify other advantages later.

Theorem 1. ***Maximal Mutual Information****. Let $\boldsymbol{MB}$ be the Markov Blanket for a target T and $\boldsymbol{V}$ be the entire feature set of the target T, $I(\boldsymbol{MB};T) \geq I(\boldsymbol{S};T), \forall \boldsymbol{S} \subseteq \boldsymbol{V}$.*

Theorem 1 states that features selected by the MB discovery methods contain all the information about the target variable. Following Theorem 1, it can be shown that the MB is the smallest feature set that contains the most mutual information to the target. Adding features to the MB feature set does not increase the mutual information and removing features from MB loses mutual information.

Theorem 2. ***Minimum Feature Set****. Let $\boldsymbol{MB}$ be the Markov Blanket for a target T and $\boldsymbol{V}$ be the entire feature set to the target T. Let $X \in \boldsymbol{V} \setminus \boldsymbol{MB}$ and $Y \in \boldsymbol{MB}$, then: 1)$I(\boldsymbol{MB};T) = I(\boldsymbol{MB} \cup X;T)$, and 2)$I(\boldsymbol{MB};T) > I(\boldsymbol{MB} \setminus Y;T)$.*

Several works [36, 46] argue that feature selection is ultimately for classification and it is ideal to choose features that minimize the classification error directly, with the Bayes error as the discriminative optimal criterion. Here we prove that using MB as the feature set also minimizes the Bayes classification error. Specifically, let $\boldsymbol{V}$ be a n-dimentional feature vector, C be the classes with a prior distribution $p(C)$ and a probability density function $p(\boldsymbol{V}|C)$, Bayes error [11] or the probability of misclassification is defined as:

$$\epsilon = 1 - \int_{\mathscr{R}^n} \max_{1 \leq i \leq C} p(i)p(\boldsymbol{V}|i)d\boldsymbol{V} \tag{1}$$

The feature selection discriminative criterion can be formulated as the following, where $\boldsymbol{S}$ indicates different feature subsets:

$$\hat{\boldsymbol{S}} = \operatorname*{argmin}_{\boldsymbol{S}} \epsilon_{\boldsymbol{S}} \tag{2}$$

Direct minimization of the Bayes error is a difficult problem, and many indirect approaches have been proposed. We choose to minimize the Bayes error by minimizing Bayes error bounds related to mutual information. Two important bounds, Fano's lower bound (LB) [9] and Hellman's upper bound (UB) [16] are particularly helpful:

$$LB : \epsilon_{\boldsymbol{S}} \geq \frac{H(T) - I(T; \boldsymbol{S}) - H(p(\epsilon))}{\log_2(C-1)} \tag{3}$$

$$UB : \epsilon_{\boldsymbol{S}} \leq \frac{H(T) - I(T; \boldsymbol{S})}{2} \tag{4}$$

where $\boldsymbol{S}$ is a feature set and C is the total number of classes in the class variable T.

Theorem 3. ***Minimal Bayes Error**. Let $\boldsymbol{MB}$ be the Markov Blanket for a target T and $\mathbf{V}$ be the entire feature set to the target T. The MB feature set minimizes both the LB and UB of $\epsilon_{\boldsymbol{S}}$.*

The formal proofs for the above three theorems can be found in the supplementary material.

4. Computer Vision Applications

In this section, we demonstrate the performance of the proposed structured feature selection for two computer vision applications: large-scale feature selection and hierarchical feature selection.

4.1. Large-Scale Feature Selection using MB

MB discovery is computationally expensive. Despite its improved efficiency over the existing methods, STMB still cannot effectively scale up to large datasets. To address this issue, we propose a new method that enables MB discovery methods to scale up with minimal or no property violations.

Inspired by feature compression methods such as PQ [17], we divide the entire feature space into d different segments with K features in each segment $\mathbf{S}_i$, with $\cup_i \mathbf{S}_i = \mathbf{V}, i = 1, ..., d$. Then we find one MB, $\mathbf{MB_i}$, of the target T within each segment $\mathbf{S}_i$. It can be shown that the PC set of T, $\mathbf{PC}_T$, will always be in $\cup_i \mathbf{MB_i}$, because $\mathbf{PC}_T$ is always dependent on the target regardless of its conditioned set and will always remain in one of $\mathbf{MB_i}$. However, the learned MB's from each segment are mostly likely not the MB learned from the entire feature space. With this procedure, the true MB variables are not necessarily present at each individual segment and learned MBs of each segment may contain false positive MB variables. This problem is solved by repeating this divide-and-conquer procedure (line 5∼6). From $\cup_i \mathbf{MB_i}$, we can break current features into a few more segments and re-discover a MB within each segment to further prune features, if the returned feature size still remains undesirably large. This procedure breaks down the large number of features and eases the computational load, while still finding the true PC set in the end. If desired, after finding MBs for each segment, left-out spouse nodes are added back in by finding features that are conditionally dependent on the target. The procedure is listed in Algorithm 2.

Algorithm 2 Large-Scale MB Feature Selection

```
 1: Input: Feature, $D$; target node, $D_T$; Folds, $d$
    {Step 1: find MB for each segment}
 2: break $D$ into $D_i$, $i \in 1, ..., d$;
 3: find MB$_i$ of $T$ within each $D_i$ using STMB;
 4: $Feat \leftarrow \cup_i MB_i$;
    {Step 2: repeat if smaller feature size is desired}
 5: $D_{Feat} \leftarrow D(:, Feat)$;
 6: Repeat Step 1 with $D_{Feat}$;
    {Step 3: add back some spouses}
 7: for each $X_i \in \mathbf{V} \setminus Feat$ do
 8:     if $X_i \not\perp\!\!\!\perp T | Feat$ then
 9:         $Feat \leftarrow Feat \cup X_i$
10: return $Feat$;
```

Algorithm 2 ensures MB is preserved with only the minimal property that could be violated. It could happen if the MB set for each segment contains false positive PC nodes. It is possible to further prune out of these undesirable variables by calling the MB discovery algorithm on the joint feature set once more.

Existing feature selection algorithms do not scale up to tens or hundreds of thousands features (even for algorithms with quadratic time complexities) [21]. The method proposed in [48] is the first work that aims to use feature selection in large-scale. Our proposed method tackles such a complexity issue by operating on a very small subset of variables and guaranteeing the optimality as if operating on the entire feature set, while all the other methods cannot guarantee so. In addition, it is unclear whether using algorithms with symmetry constraints such as IPCMB with Algorithm 2 would succeed to find all PC nodes of the target. PC of neighbor nodes of a target on a subset of features can be very different from those on the entire feature set and thus could potentially remove the correct PC nodes. Therefore, IPCMB cannot guarantee to find the correct MB with Algorithm 2, while STMB does not have this problem.

We empirically compare the performance of our structured feature selection in Algorithm 2 to existing feature selection and feature compression methods. We use the popular VOC 2007 dataset [8] and extract Fisher Vector on dense SIFT features from 8 spatial pyramid regions, following existing works [5, 48]. To handle the large feature size s of FV (s = 262144), we employ Algorithm 2 with mutual-information-based independence tests. The tests check if the (conditional) mutual information between two variables is bigger than a predefine significance level; if so, these two variables are dependent, otherwise independent. In addition, to quantize continuous FV features, we use a 1-BIT quantization to calculate the mutual information [48]: a real number x becomes 0 if $x <= 0$, or 1 if $x > 0$. The proposed STMB has the innate compression ratio of 32 due to the 1-BIT quantization. We break down the features into segments of sizes K, and compare the performance of the selected features with existing methods [42, 48]. The results are shown in Table 1. The compression ratio is calculated as $32 \cdot \frac{s}{m}$, where m is the average feature size over all the object categories. Different compression ratios in Table 1 indicate different numbers of selected features. The proposed MB-based method achieves -5% relative mAP (to the baseline) at the compression ratio of 608, while the MI-based method [48] achieves -5.87% relative performance at the compression ratio of 512 and PQ achieves -4.8% relative performance at the compression ratio of 128. Moreover, Figure 2 shows that some categories only need a few features to achieve similar performance. For example, Category 5 needs 2320 features for the compression ratio 608 to maintain a -6.05% relative performance, which has a category compression ratio of 3616! The results show our method can select a smaller set at high compression ratios with less loss than other approaches. We believe with a better independence test our method can perform even better.

In addition, results from Table 1 show that different K sizes have minimal impact on the final accuracy. This indicates that more speed-ups can be achieved by dividing data into more segments and using more threads simultaneously with similar performance. It is also worth noting that although breaking data into single elements in each segment is the most efficient, but they would save many more false positive nodes.

Table 1. mAP and Feature Size on VOC 2007 with linear SVM

Method	Compression Ratio	mAP (%)	mean Feature Size
MB	1	59.4	262144
	608, $K = 100$	54.4	13415
	704, $K = 200$	54.0	11405
MI [48]	1	58.57	262144
	256	56.82	32768
	512	52.70	16384
	1024	46.52	8192
PQ [42]	1	58.8	262144
	128 ($d = 8$)	54.0	262144
	256 ($d = 8$)	50.3	262144

4.2. Hierarchical Feature Selection

In this section, we demonstrate the application of structured feature selection for hierarchical feature selection, i.e., features with different levels of abstractions. Specifically, we apply our method to feature selection for facial expression and action unit (AU) recognition. For facial expression recognition, features can be divided into shape features, e.g., facial feature points (FFP), and local facial motions, e.g., facial AUs. They form a top down hierarchy as Expression → AUs → FFPs . Selected FFPs can be used for AU recognition, and selected FFPs and AUs can be used for expression recognition. We apply the proposed method to select features for both AU and expression recognition.

Features. We use the popular CK+ face dataset [24], which consists of 593 video sequences from 123 subjects. We use the peak frame data for the expression and the most popular 15 AU labels in each video sequence. We use the provided shape features only (i.e., FFPs) to predict AUs and expressions, and do not use any appearance or dynamic features. We normalize faces to a fixed facial model by interocular distance [38] and use the normalized frontal 51 FFPs as features. We further quantize the FFPs into 4 directions, by their relative positions to neutral positions (i.e., peak frame positions versus the first frame positions). All the procedure follows standard protocols [45, 23, 22].

Methodology. We first find the MBs using STMB for the expression with respect to all 15 AUs and 51 FFPs (i.e., the target variable is the expression), and for each AU with respect to the expression and FFPs (i.e., the target variable is each AU). To fully capture the hierarchical structure information, we build a global hierarchical BN from the learned MBs. To preserve the independence relationships among variables, we first orient all the V-structures and use the Meek rules[2] [25] to orient as the rest of links as possible. Then any unresolved edges are oriented as Exp → AU → FFP to lower the number of parameters, thus reducing the

[2] The details of Meek rules can be found in the supplementary material.

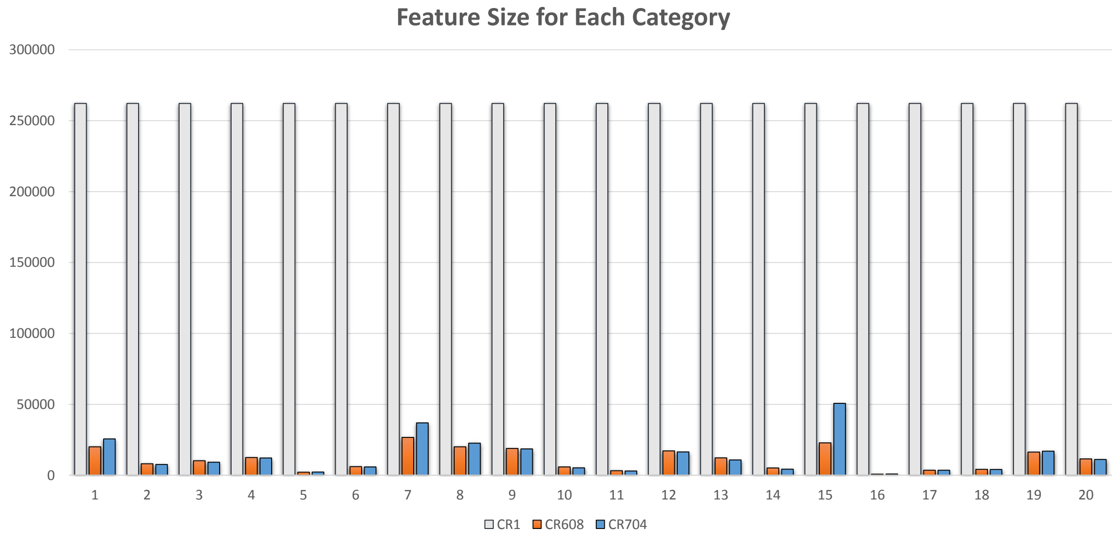

Figure 2. Selected feature sizes for each category in VOC 2007 by our proposed method, with different compression ratios (CRs), corresponding to results from Table 1.

graph complexity and data requirement for accurate learning. This local-to-global hierarchical BN structure learning procedure is summarized in Algorithm 3. Note that directly learning such a BN is not feasible. BN structure learning algorithms generally require $O(2^N)$ memory [39]. They cannot handle the 67 nodes in the CK+ dataset, which would need 10^{11} GB memory. MB discovery can be seen as a subproblem of BN structure learning, and can be used to build BN (as seen by Algorithm 3), which enables BN structure learning to handle much large sizes of variables with a divide-and-conquer approach.

Algorithm 3 Local-to-Global BN Structure Learning

```
Input: Data, $D$; total variables, $V$
{Step 1: find MB for emotion and each AU}
for each $X_i \in \mathbf{AU} \cup \mathbf{Exp}$, do
    MB$_i \leftarrow$ STMB($D$, $X_i$);
{Step 2: combine local structures}
$DAG \leftarrow zeros(|V|, |V|)$;
if $X_m \in MB_n, \forall m \in V, n \in \mathbf{AU} \cup \mathbf{Exp}$,, then
    $DAG(n, m) = 2$; //undirected
if $DAG(m, n) \neq DAG(n, m), \forall m \in V, n \in \mathbf{AU} \cup \mathbf{Exp}$, then
    $DAG(m, n) \leftarrow 0$;
    $DAG(n, m) \leftarrow 0$; //enforce symmetry
if $n \in Spouse_m, m \in Spouse_n, \forall m \in V, n \in \mathbf{AU} \cup \mathbf{Exp}$, then
    find the common children $Z$;
    $DAG(n, Z) = 1$;
    $DAG(m, Z) = 1$; //V-structure
{Step 3: Resolve edge orientation}
enforce Meek Rule [25] on $DAG$;
if $DAG(m, n) = 2$, then
    orient $DAG(m, n)$ following $Exp \rightarrow AU \rightarrow FFP$;
return $DAG$;
```

We only find the MBs of the expression and AUs, and enforce the symmetry constraint only during the local-to-global hierarchical structure learning process but not during the local discovery. It is mainly because finding the MBs of the FFPs, which have very dense local structures, takes a very long time. MB algorithms with the symmetry constraint would need to find MBs for these FFPs, which may take much longer time and limit the practical usage. We divide data into five subject-independents [38] folds, and learn the global hierarchical structure and parameters on four folds and test on the other. Figure 3 shows the learned hierarchical structures on two different folds. Ideally the learned structures should be the same across all folds, but the pure data-driven approach can learn different structures to capture the joint distribution of the variables in a BN. During testing, we use the MAP inference, based on the junction tree method, to infer AUs with selected FFPs as evidence. We then use the inferred values of AUs and FFPs jointly as evidence to infer the expression with MAP.

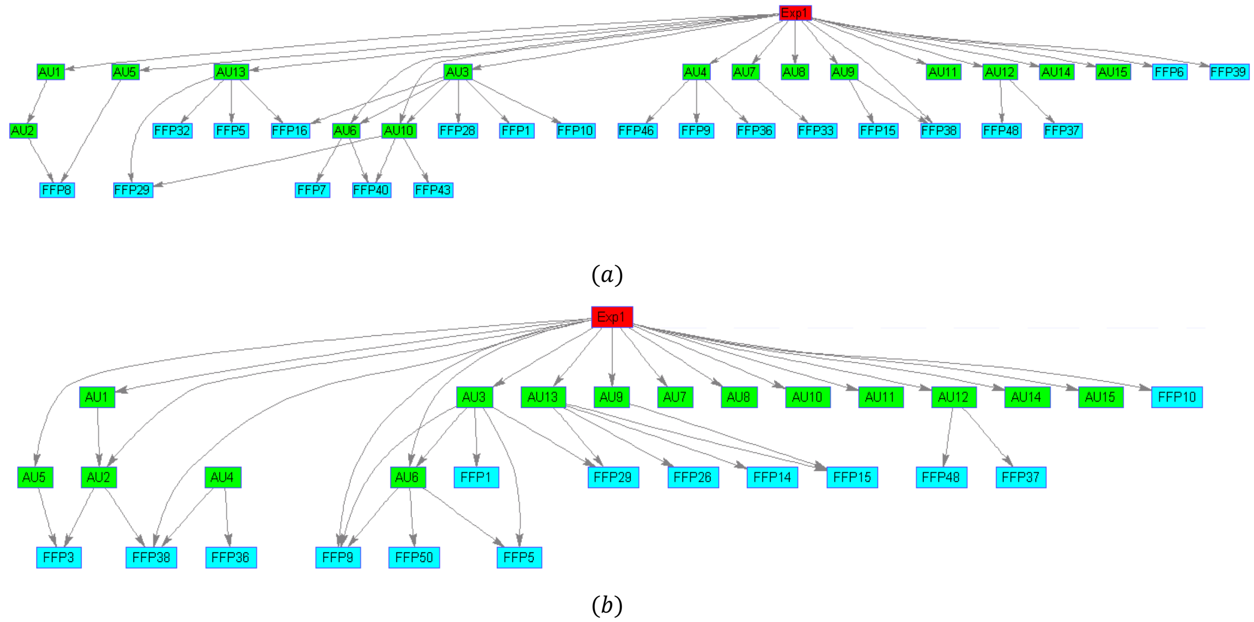

Figure 3. Learned Global Structures of CK+ dataset for 5-fold Cross Validation. a) is learned from fold 2 to 5, and b) is learned from fold 1, 3,4, and 5.

4.2.1 AU Recognition Experiment Results

First we show the performance for AU recognition using the learned hierarchical structure. We use F_1-score to measure the accuracy and results are shown in Table 2. Using the selected FFPs from STMB leads to a 6% improvement over using all the FFPs with SVM, and our proposed local-to-global method has a performance increase of about 16% over SVM with the same selected FFPs. We also compare our feature selection method with some standard feature selection methods, such as ReliefF, Lasso, and mRMR [32]. We test these algorithms with an SVM classifier and the BN constructed using Algorithm 3. Although selected features from STMB are not the best with SVM, the proposed method with BN outperforms all other feature selection and classifier combinations.

The results from Table 2 also show that, despite our simple features and less training data with fewer fold numbers, the AU recognition rate on a smaller set of AUs (15 vs. 17) is very close to the previous state-of-the-art BN result [40]. If we consider the expression as a feature, our method can lead to an even bigger improvement on smaller folds' cross-validation, as seen in the bottom part of Table 2, on smaller folds cross validation.

Table 2. AU Recognition, Average F_1-score, on CK+

Method	Fold	Features	F_1-score
SVM	5	All FFP	52.42 %
BN-Lasso	5	Selected FFP	55.38 %
SVM-RELIEFF	5	Selected FFP	56.64 %
SVM-STMB	5	Selected FFP	58.67 %
SVM-Lasso	5	Selected FFP	58.97 %
SVM-MRMR	5	Selected FFP	67.36 %
BN-MRMR	5	Selected FFP	70.91 %
BN-STMB	5	Selected FFP	74.52 %
BN [40, 44]	LOSO	FFP & Appearance	76.70%
HRMB[44]	LOSO	FFP & Appearance	82.44%
SVM	2	All FFP & Exp	70.64 %
SVM-STMB	2	Selected FFP & Exp	68.18%
BN-STMB	2	Selected FFP & Exp	86.68%

4.2.2 Expression Recognition Experiment Results

We train an SVM using the FFP features as the baseline method and compare the Average Recognition Rate (ARR), calculated from the diagonal of the confusion matrix, for different methods. The results are shown in Table 3. Using the selected FFPs (of average size 22.3) has a lower ARR compared to using all the FFPs (of size 51) with SVM, but our proposed hierarchical method uses less than half of the total FFPs and inferred AUs, and outperforms baseline SVM methods by about 3%. SVM may not perform well when selected features are not optimal. This is, however, less a problem for BN since it can still use the structural relationships among selected features. Compared to other feature selection algorithms, the proposed approach also gives the best ARR overall. Moreover, we test the expression recognition using ground truth AUs in addition to FFPs as features, shown in the bottom part of Table 3. We only use 2 folds, which contains less training data compared to the standard 15-fold or even leave-one-subject-out (LOSO) cross validation. This shows the upper bound for the expression recognition performance using the learned BN structures. It validates that the learned hierarchical structures can capture the correct relationships among AUs and FFPs, and the proposed method can lead to significant improvement, potentially with better AU estimation from a richer set of features.

Table 3. Expression Recognition Rate on CK+

Method	Fold	Features	ARR
BN-LASSO	5	Selected FFP	55.29 %
SVM-RELIEFF	5	Selected FFP	59.75 %
SVM-STMB	5	Selected FFP	63.43 %
SVM-LASSO	5	Selected FFP	64.02 %
SVM-MRMR	5	Selected FFP	65.29 %
SVM	5	All FFP	69.50 %
BN-MRMR	5	Selected FFP	69.89%
BN-STMB	5	Selected FFP	72.28 %
Lucey et al [24]	LOSO	Appearance & Shape	83.3%
ITBN [45]	15	Dynamic & FFP	86.3%
SVM	2	All FFP & AU	89.59 %
SVM-STMB	2	Selected FFP & AU	91.19%
BN-STMB	2	Selected FFP & AU	96.81 %

Note that the proposed method has 56% and 50% recognition rates on "Fear" and "Sadness", two expressions that have a very small sample size (25 and 28 respectively), while achieving a mean ARR of 82% on the other 4 expressions. Since our approach is purely data driven, it is more sensitive to sample sizes than the knowledge-driven approaches [45].

5. Discussion and Conclusion

This paper proposes to use MB-based feature selection for dimensionality reduction in computer vision tasks, and contains four major contributions. First, we propose a new MB discovery algorithm, improving the efficiency of existing algorithms. Secondly, we further enrich and prove the theoretical advantages of structured feature selection using MB discovery compared to traditional feature selection approaches. Thirdly, we propose a scaling-up method for MB-based feature selection in large-scale feature selection tasks and apply it to image classification. Lastly, in facial expression recognition tasks, we show the competitive performance of the MB-based feature selection method for hierarchical features. In particular, for high feature compression ratios, the proposed method can achieve higher accuracy than state-of-the-art methods with a lower number of features. The structures that the proposed hierarchical feature selection method discovers can also lead to big performance gain in recognition tasks.

It is worth noting that, compared to existing MB discovery algorithms, STMB does not use the symmetry constraint, which improves the efficiency of the MB discovery by $O(|\mathbf{PC}|)$. In addition, it is unclear whether other MB algorithms with the symmetry constraint can guarantee to find all MB nodes of the target with the proposed large scale feature selection algorithm. Moreover, for some applications, the symmetry constraint would need to find MBs for non-target or trivial nodes (such as FFPs in hierarchical feature selection), whose MB sizes can be very large, limiting the practical usage of MB discovery algorithms. For example, IPCMB cannot be directly applied to large computer vision datasets because of the time complexity with its symmetry check step, which can take up to 262144/200 = 1311 ($K = 200$) more times than STMB in VOC07 and 67 more times in the CK+ dataset.

The experimental results also show that structures among features can provide additional information to increase classification accuracy, and it would be an interesting direction to study methods of incorporating these structure information into traditional classifiers like SVM systematically. For example, one intuitive approach is to encode the structure information (such as features' identities with respect to each other and the target in terms of parents, children, or neither) as additional features. We would also like to test this proposed method on other large-scale computer vision datasets and on other tasks like human gesture recognition, which contains similarly rich structure information.

References

[1] M. R. Abid, E. M. Petriu, and E. Amjadian. Dynamic sign language recognition for smart home interactive application using stochastic linear formal grammar. *Instrumentation and Measurement, IEEE Transactions on*, PP(99), 2014. 2

[2] C. Aliferis, I. Tsamardinos, A. Statnikov, C. F. Aliferis, I. Tsamardinos, and E. Statnikov. Hiton, a novel markov blanket algorithm for optimal variable selection, 2003. 2

[3] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *JMLR*, Jan 2010. 2, 3

[4] A. Bulling, J. A. Ward, H. Gellersen, and G. Troster. Eye movement analysis for activity recognition using electrooculography. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(4):741–753, 2011. 1

[5] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011. 1, 5

[6] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *CVPR*, pages 3025–3032. IEEE, 2013. 1

[7] M. Das Gupta and J. Xiao. Non-negative matrix factorization as a feature selection tool for maximum margin classifiers. In *CVPR*, pages 2841–2848. IEEE, 2011. 1

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 5

[9] R. M. Fano. *Transmission of Information: A Statistical Theory of Communication*. MIT Press, MA, USA, 1961. 4

[10] S. Fu and M. C. Desmarais. Fast markov blanket discovery algorithm via local learning within single pass. In *Proceedings of the Canadian Society for computational studies of intelligence, 21st conference on Advances in artificial intelligence*, Canadian AI'08, pages 96–107, Berlin, Heidelberg, 2008. Springer-Verlag. 2, 3

[11] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 1990. 4

[12] T. Ge, K. He, Q. Ke, and J. Sun. Optimized product quantization for approximate nearest neighbor search. In *CVPR*, pages 2946–2953. IEEE, 2013. 1

[13] Y. Gong, S. Kumar, H. A. Rowley, and S. Lazebnik. Learning binary codes for high-dimensional data using bilinear projections. In *CVPR*, pages 484–491. IEEE, 2013. 1

[14] B. Han, M. Park, and X.-w. Chen. A markov blanket-based method for detecting causal snps in gwas. *BMC bioinformatics*, 11(Suppl 3):S5, 2010. 3

[15] R. He, T. Tan, L. Wang, and W.-S. Zheng. l 2, 1 regularized correntropy for robust feature selection. In *CVPR*, pages 2504–2511. IEEE, 2012. 1

[16] M. E. HELLMAN and J. RAVIV. Probability of error, equivocation, and the. *IEEE Transactions on Information Theory*, 16(4), 1970. 4

[17] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):117–128, 2011. 1, 4

[18] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1704–1716, 2012. 1

[19] D. Koller and M. Sahami. Toward optimal feature selection. In *ICML 1996*, pages 284–292. Morgan Kaufmann, 1996. 2, 3

[20] D. Kong, C. Ding, H. Huang, and H. Zhao. Multi-label relieff and f-statistic feature selections for image annotation. In *CVPR*, pages 2352–2359. IEEE, 2012. 1

[21] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):491–502, 2005. 4

[22] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *CVPR*, pages 1749–1756. IEEE, 2014. 5

[23] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *CVPR*, pages 1805–1812. IEEE, 2014. 5

[24] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*, pages 94–101. IEEE, 2010. 5, 8

[25] C. Meek. Causal inference and causal explanation with background knowledge. In *UAI*, pages 403–410. Morgan Kaufmann Publishers Inc., 1995. 5, 6

[26] N. Naikal, A. Yang, and S. Sastry. Informative feature selection for object recognition via sparse pca. In *ICCV*, pages 818–825, Nov 2011. 1

[27] M. Norouzi and D. J. Fleet. Cartesian k-means. In *CVPR*, pages 3017–3024. IEEE, 2013. 1

[28] T.-H. Oh, H. Kim, Y.-W. Tai, J.-C. Bazin, and I. S. Kweon. Partial sum minimization of singular values in rpca for low-level vision. In *ICCV*, pages 145–152. IEEE, 2013. 1

[29] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers, Inc., 2 edition, 1988. 2

[30] J. Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000. 3

[31] J. M. Peña, R. Nilsson, J. Björkegren, and J. Tegnér. Towards scalable and data efficient learning of markov boundaries. *Int. J. Approx. Reasoning*, 45(2):211–232, July 2007. 2, 3

[32] H. Peng, F. Long, and C. Ding. Feature selectin based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Learning*, 27(8):1226–1238, 2005. 7

[33] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, pages 1–8. IEEE, 2007. 1

[34] M. J. Saberian and N. Vasconcelos. Boosting algorithms for simultaneous feature extraction and selection. In *CVPR*, pages 2448–2455. IEEE, 2012. 1

[35] J. Sánchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In *CVPR*, pages 1665–1672. IEEE, 2011. 1

[36] G. Saon and M. Padmanabhan. Minimum bayes error feature selection. In *INTERSPEECH*, pages 75–78, 2000. 4

[37] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human detection using partial least squares analysis. In *ICCV*, pages 24–31. IEEE, 2009. 1

[38] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009. 5, 6

[39] T. Silander and P. Myllymaki. A simple approach for finding the globally optimal bayesian network structure. In *UAI*, pages 445–452, 2006. 6

[40] Y. Tong and Q. Ji. Learning bayesian networks with qualitative constraints. In *CVPR*, pages 1–8. IEEE, 2008. 7

[41] I. Tsamardinos, C. Aliferis, A. Statnikov, and E. Statnikov. Algorithms for large scale markov blanket discovery. In *FLAIRS*, pages 376–380. AAAI Press, 2003. 2

[42] A. Vedaldi and A. Zisserman. Sparse kernel approximations for efficient classification and detection. In *CVPR*, pages 2320–2327. IEEE, 2012. 5

[43] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *ICCV*, pages 2088–2095. IEEE, 2013. 1

[44] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *ICCV*, pages 3304–3311. IEEE, 2013. 7

[45] Z. Wang, S. Wang, and Q. Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *CVPR*, pages 3422–3429. IEEE, 2013. 5, 8

[46] S. H. Yang and B.-G. Hu. Discriminative feature selection by nonparametric bayes error minimization. *Knowledge and Data Engineering, IEEE Transactions on*, 24(8):1422–1434, 2012. 4

[47] X. Zeng, D. F. Wong, L. S. Chao, I. Trancoso, L. He, and Q. Huang. Lexicon expansion for latent variable grammars. *Pattern Recognition Letters*, 42:47–55, 2014. 2

[48] Y. Zhang, J. Wu, and J. Cai. Compact representation for image classification: To choose or to compress? In *CVPR*. IEEE, 2014. 1, 4, 5

[49] X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, pages 141–154. Springer, 2010. 1

[50] X. Zhu, H.-I. Suk, and D. Shen. Matrix-similarity based loss function and feature selection for alzheimer's disease diagnosis. In *CVPR*, pages 3089–3096. IEEE, 2014. 1