Feature Selection Using Correlation and Reliability Based Scoring Metric for Video Semantic Detection

Qiusha Zhu, Lin Lin, Mei-Ling Shyu
Department of Electrical and
Computer Engineering
University of Miami
Coral Gables, FL 33124, USA
Email: {q.zhu2, l.lin2}@umiami.edu, shyu@miami.edu

Shu-Ching Chen
School of Computing and
Information Sciences
Florida International University
Miami, FL 33199, USA
Email: chens@cs.fiu.edu

***Abstract*—Content-based multimedia retrieval faces many challenges such as semantic gap, imbalanced data, and varied qualities of the media. Feature selection as a component of the retrieval process plays an important role. The aim of feature selection is to identify a subset of features by removing irrelevant or redundant features. An effective subset of features can not only improve model performance and reduce computational complexity, but also enhance semantic interpretability. To achieve these objectives, in this paper, a novel metric that integrates the correlation and reliability information between each feature and each class obtained from Multiple Correspondence Analysis (MCA) is proposed to score the features for feature selection. Based on these scores, a ranked list of features can be generated and different selection criteria can be adopted to select a subset of features. To evaluate the proposed framework, four other well-known feature selection methods, namely information gain, chi-square measure, correlation-based feature selection, and relief are compared with the proposed method over five popular classifiers using the benchmark data from TRECVID 2009 high-level feature extraction task. The results show that the proposed method outperforms the other methods in terms of classification accuracy, the size of feature subspace, and the ability to capture the semantic information.**

I. Introduction

With the increasing number of high-dimensional datasets ranging from several hundred to hundred thousand features, the process of selecting a good feature subset has become more and more important. Such a process can remove irrelevant, redundant, or noisy features to improve model performance and make a model more cost-effective. In addition to these objectives, feature selection can also preserve the semantics of the features compared to other dimensionality reduction techniques. Instead of altering the original representation of features like those based on projection (e.g., principal component analysis) and compression (e.g., information theory) [1], feature selection eliminates those features with little predictive information, keeps those with better representation of the underlying data structure, and thus enhances the semantic interpretability.

In recent years, different areas have adopted the feature selection technique to pre-process the data in order to improve model performance. In general data mining and pattern recognition domains, Liu and et al. [2] introduced a criterion function of mutual information and proposed a mutual information based feature selection method which could generate a subset of features without taking class labels into account. Experimental results on 16 commonly used UCI datasets using four typical classifiers showed that the proposed method outperformed relief, information gain, and symmetric uncertainty (SU) in most cases. However, their proposed method was time consuming since the values of mutual information of a feature subset needed to be re-calculated after a feature had been chosen, and it was also sensitive to noise. In [3], the comparison of some famous feature selection methods in the area of bioinformatics was given, including information gain, gini index, t-test, sequential forward selection (SFS), and etc. Feature selection in this area is inevitable but quite challenging because biological technologies usually produce data with thousands of features but a relatively small sample size. Experiments on both synthetic and real data showed that none of these methods performed best across all scenarios, and revealed some trend relative to the sample size and relations among the features.

As an important application area, content-based multimedia retrieval has become a very popular research direction since the amount of multimedia data keeps growing exponentially nowadays. A great deal of efforts have been dedicated to challenging topics in this field, such as bridging the semantic gap between low-level features and high-level concepts, handling highly imbalanced ratios between positive and negative instances, and ensuring robust performance on media with varied qualities. Being a vital processing step, feature selection can reduce the cost of storage, decrease redundancy, and improve the performance of the model in these aspects. An effective subset of features should not contain (i) noisy features that decrease the retrieval accuracy, or (ii) irrelevant features that increase the computation time. Instead, it should contain those that have high predictive information and could better capture the semantic meaning of the query sample. Thus, a good feature selection can intrinsically help content-based retrieval overcome these challenges.

Since multimedia data are numeric, as a straightforward approach, PCA (Principal Component Analysis) has been used as a feature selection method in many literatures [4]. It alters

the original representation of the features by projecting all of them into a low dimensional space and combining them linearly. MCA (Multiple Correspondence Analysis) [5] on the other hand, is designed for nominal data. However, if it can be effectively utilized to indicate the relations between a feature and a class, MCA could be considered as a potentially better approach since by choosing a subset from the original feature space, the semantic meaning of the feature is retained. In our previous studies [6] [7], the angle values obtained from MCA have shown to be able to capture the correlation between each feature and the class. However, according to [8], correlation may not be sufficient and accurate enough to describe the data structure, and thus it studied complex feature dependencies in multivariate settings for multimedia information fusion. The proposed approach in this paper, from the view of univariate settings, continues to explore the geometrical representation of MCA and aims to find an effective way to indicate the relation between features and classes. In statistics, p-value serves as a measure of reliability of a relation. This motivates us to utilize p-values as a measure of reliability of the relation between a feature and a class. The contribution of this paper is proposing a novel feature scoring metric that integrates both correlation and reliability information represented by the angle values and p-values. Based on the metric, a score is calculated for each feature and a feature subset can be selected according to the ranking scores. Please note that the focus is on the problem of supervised learning, which means that the classification is carried out with the labels of the training dataset known beforehand.

To evaluate our proposed scoring metric, four well-known feature selection methods for supervised learning, namely the information gain, chi-square measure, correlation-based feature selection, and relief filter methods, available in WEKA [9] are used in the performance comparison using Decision Tree , Rule based JRip, Native Bayes, Adaptive Boosting, and K-Nearest Neighbor classifiers. More detailed descriptions of these classifiers could be found in [9]. The data used in the experiments is from TRECVID 2009 high-level feature extraction task, where the high-level features are actually the concepts to be detected in video shots. The overall experimental results demonstrate that the proposed scoring metric which integrates angle values and p-values performs better than other feature selection methods over all these five classifiers, not only in classification accuracy and the number of dimensions reduced, but also in the ability to capture the semantic information of the video concepts.

The paper is organized as follows. Related work is introduced in Section II. Our proposed framework is presented in Section III, followed by an analysis of the experimental results in Section IV. Finally, conclusions are given in Section V.

II. Related Work

Depending on how it is combined with the construction of the classification model, supervised feature selection can be further divided into three categories: wrapper methods, embedded methods, and filter methods. Wrappers choose feature subsets with high prediction performance estimated by a specified learning algorithm which acts as a black box, and thus wrappers are often criticized for their massive amounts of computation which are not necessary. Similar to wrappers, embedded methods incorporate feature selection into the process of training for a given learning algorithm, and thus they have the advantage of interacting with the classification model, meanwhile being less computationally intensive than wrappers. These two categories usually yield better classification results than the filter methods, since they are tailored to a specific classifier, but the improvements of the performance are not always significant because of the curse of dimensionality and the fact that the specific tuned classifiers may overfit the data. In contrast, filter methods are independent of the classifiers and can be scaled for high-dimensional datasets while remaining computationally efficient. In addition, filtering can be used as a pre-processing step to reduce space dimensionality and overcome the overfitting problem. Therefore, filter methods only need to be executed once, and then different classifiers can be evaluated based on the generated feature subsets.

Filter methods can be further divided two main sub-categories. The first one is univariate methods which consider each feature with the class separately and ignore the inter-dependence between the features. Representative methods in this category include information gain and chi-square measure, both of which are widely used to measure the dependence of two random variables. Information gain evaluates the importance of features by calculating their information gain with the class, but this method is biased to features with more values. In [10], a new feature selection method was proposed which selected features according to a combined criterion of information gain and novelty of information. This criterion strives to reduce the redundancy between features while maintaining information gain in selecting appropriate features. In contrast, chi-square measure calculates the χ^2 statistics between each feature and the class, and a large value indicates a strong correlation between them. Although this method does not adhere strictly to the statistics theory because the probability of errors increases when a statistical test is used multiple times, it is applicable as long as it only ranks features with respect to their usefulness [11]. Jiang et al. [12] used the bag-of-visual-words (BoW) features to represent keypoints in images for semantic concept detection. As one of the representation choices of BoW, feature selection applied the chi-square measure to calculate the χ^2 statistics between a specific visual word and a binary label of an image class, and eliminated those virtual words with χ^2 statistics below a threshold. Extensive experiments on the TRECVID data indicated that BoW features with appropriate representation choices could produce highly competitive results.

The second sub-category is the multivariate methods which take features' interdependence into account. However, they are slower and less-scalable compared to the univariate methods. Correlation-based feature selection (CFS) is one of the most popular methods. It searches among the features according to the degree of redundancy between them in order to find a

subset of features that are highly correlated with the class, yet uncorrelated with each other [13]. Experiments on natural datasets showed that CFS typically eliminated over half of the features, and the classification accuracy using the reduced feature set was usually equal to or better than the accuracy using the complete feature set. The disadvantage is that CFS degrades the performance of classifiers in cases where some eliminated features are highly predictive of very small areas of the instance space. This kind of cases could be frequently encountered when dealing with imbalanced data. Relief is another commonly used method whose idea is to choose the features that can be most distinguishable between classes. It evaluates the worth of a feature by repeatedly sampling an instance and considering the value of the given feature for the nearest instance of the same and different classes. However, relief lacks a mechanism to deal with the outlier instances, and according to [3], it has worse performance than the univariate filter methods in most cases. Sun [14] proposed an iterative relief (I-Relief) method by exploring the framework of the Expectation-Maximization algorithm. Large-scale experiments conducted on nine UCI datasets and six microarray datasets demonstrated that I-Relief performed better than relief without introducing a large increase in computational complexity.

According to the form of the outputs, the four aforementioned feature selection methods can also be categorized into ranker and non-ranker. A non-ranker method provides a subset of features automatically without giving an order of the selected features such as CFS. On the other hand, a ranker method provides a ranked list by scoring the features based on a certain metric, to which information gain, chi-square measure, and relief belong. Then different stopping criteria can be applied in order to get a subset from it. Most commonly used criteria include forward selection, backward elimination, bi-directional search, setting a threshold, genetic search, etc.

III. The Proposed Framework

Our proposed framework utilizes the correlation and reliability information from MCA and integrates them into a single feature scoring metric to evaluate features.

A. Geometrical Representation of MCA

Standard Correspondence Analysis (CA) is a descriptive/exploratory technique designed to analyze simple two-way contingency tables containing some measure of correspondence between the rows and columns. Multiple Correspondence Analysis (MCA) can be considered as an extension of the standard CA to more than two variables [5]. Meanwhile, it also appears to be the counter part of PCA for nominal (categorical) data.

MCA constructs an indicator matrix with instances as rows and categories of valuables as columns. Here in order to apply MCA, each feature needs to be first discretized into several intervals or nominal values (called feature-value pairs in our study), and then each feature is combined with the class to form an indicator matrix. Assuming the kth feature has j_k feature-value pairs and the number of classes is m, then the indicator matrix is denoted by Z with size $n \times (j_k + m)$, where n is the number of instances. Instead of performing on the indicator matrix which is often vary large, MCA analyzes the inner product of this indicator matrix, i.e., Z^TZ, called the Burt Table which is symmetric with size $(j_k + m) \times (j_k + m)$. The grand total of the Burt Table is the number of instances which is n, then $P = Z^T Z/n$ is called the correspondence matrix with each element denoted as p_{ij}. Let r_i and c_j be the row and column masses of P, that is, $r_i = \sum_j p_{ij}$ and $c_j = \sum_i p_{ij}$. The center involves calculating the differences $(p_{ij} - r_i c_j)$ between the observed and expected relative frequencies, and normalization involves dividing these differences by $\sqrt{r_i c_j}$, leading to a matrix of standardized residuals $s_{ij} = (p_{ij} - r_i c_j)/\sqrt{r_i c_j}$. The matrix notation of this equation is presented in Equation (1).

$$S = D_r^{-1/2}(P - rc^T)D_c^{-1/2}, \tag{1}$$

where r and c are vectors of row and column masses, and D_r and D_c are diagonal matrices with these masses on the respective diagonals. Through Singular Value Decomposition (SVD), $S = U\Sigma V^T$ where Σ is the diagonal matrix with singular values, the columns of U are called left singular vectors, and those of V are called right singular vectors. The connection of the eigenvalue decomposition and SVD can be seen through the transformation in Equation (2).

$$SS^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T = U\Lambda U^T. \tag{2}$$

Here, $\Lambda = \Sigma^2$ is the diagonal matrix of the eigenvalues, which is also called principal inertia. Thus, the summation of each principal inertia is the total inertia which is also the amount that quantifies the total variance of S. The geometrical way to interpret the total inertia is that it is the weighted sum of squares of principal coordinates in the full S-dimensional space, which is equal to the weighted sum of squared distances of the column or row profiles to the average profile. This motivates us to explore the distance between feature-value pairs and classes represented by rows of principal coordinates in the full space. Since over 95% of the total variance can be captured by the first two principal coordinates [5], the χ^2 distance between a feature-value pair and a class can be well represented by the Euclidean distance between them in the first two dimensions of their principal coordinates. Thus, a graphical representation, called the symmetric map, can visualize a feature-value pair and a class as two points in the two dimensional map.

As shown in Fig. 1, a nominal feature with three feature-value pairs corresponds to three points in the map, namely $P1$, $P2$, and $P3$, respectively. Considering a binary class, it is represented by two points lying in the x-axis, where $C1$ is the positive class and $C2$ is the negative class. Take $P1$ as an example. The angle between $P1$ and $C1$ is a_{11}, and the distance between them is d_{11}. Similar to standard CA, the meaning of a_{11} and d_{11} in MCA can be interpreted as follows.

- Correlation: This is the cosine value of the angle between a feature-value pair and a class in the symmetric map. It

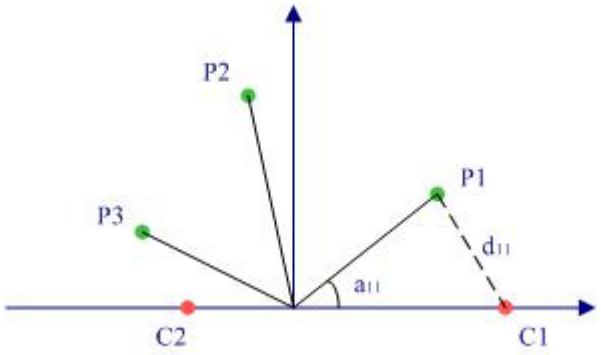

Fig. 1. The symmetric map of the first two dimensions

represents the percentage of the variance that the feature-value pair point is explained by the class point. A larger cosine value which is equal to a smaller angle indicates a higher quality of representation.

- Reliability: As stated before, χ^2 distance could be used to measure the dependence between a feature-value pair point and a class point. Here, a derived value from χ^2 distance called the p-value is used because it is a standard measure of the reliability of a relation, and a smaller p-value indicates a higher level of reliability.

Assume that the null hypothesis H_0 is true. Generally, one rejects the null hypothesis if the p-value is smaller than or equal to the significance level, which means the smaller the p-value, the higher possibility of the correlation between a feature-value pair and a class is true. P-value can be calculated through the χ^2 Cumulative Distribution Function (CDF) and the degree of freedom is (number of feature-value pairs -1) $\times$ (number of classes -1). For example, the χ^2 distance between $P1$ and $C1$ is d_{11} and their degree of freedom is $(3-1) \times (2-1)$, and then their p-value is $1-$CDF(d_{11},2). Therefore, correlation and reliability are from different points of view, and can be integrated together to represent the relation between a feature and a class.

B. MCA-based Feature Selection Framework

The proposed MCA-based feature selection framework is shown in Fig. 2 which contains three stages: MCA calculation, feature evaluation, and stopping criteria.

First, MCA is applied to nominal feature values and classes. Since the original features extracted from videos are numeric, discretization is needed before applying MCA, so each feature would be discretized into multiple feature-value pairs. For each feature, the angles and p-values between each feature-value pair of this feature to the positive and negative classes are calculated, corresponding to correlation and reliability, respectively. If the angle of a feature-value pair with the positive class is less than 90 degrees, it indicates this feature-value pair is more closely related to the positive class than to the negative class, or vice versa. For p-value, since a smaller p-value indicates a higher reliability, (1 - p-value) can be used as the probability of a correlation being true, except for

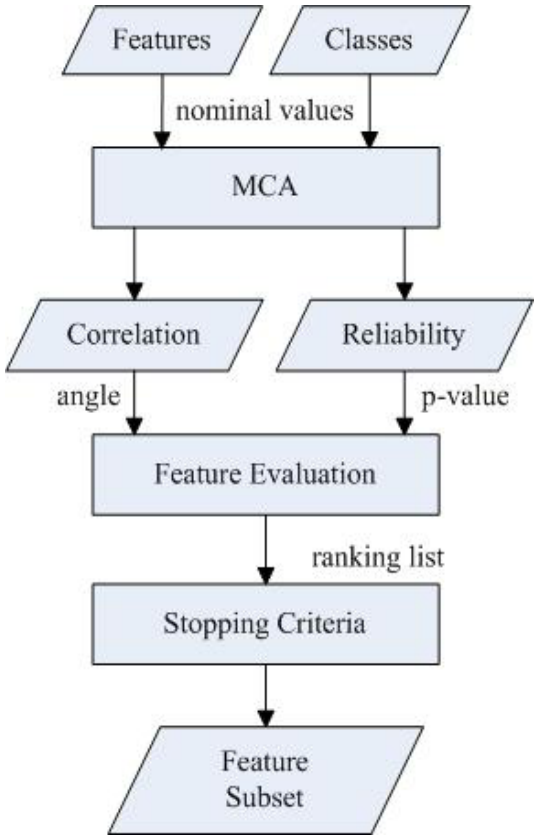

Fig. 2. The proposed feature selection framework

the situation of Jeffrey-Lindley paradox [15]. This paradox describes a situation when the p-value is very close to zero but the probability of the correlation being true is very close to zero as well. Such scenario could happen when the prior distribution is the sum of a sharp peak at H_0 with probability p and a broad distribution with the rest of the probability $1-p$. In our experiments (to be discussed in Section IV), it occurs when the count of the cross-table constructed by the feature-value pairs and the classes is less than 1% of the count of the corresponding class, which also makes sense since a rare occurrence can be considered as a "fluke".

After getting the correlation and reliability information of each feature-value pair with the class in the MCA calculation stage (represented by the angle values and p-values correspondingly), the equations which take the cosine value of an angle and p-value as two parameters are defined (as presented in Equations (3) and (4)) in the feature evaluation stage. Since these two parameters may play different roles in different datasets and both of them lie between $[0, 1]$, different weights can be assigned to these two parameters in order to sum them together as an integrated feature scoring metric. Considering different nominal features contain a different number of feature-value pairs, to avoid being biased to features with more categories like Information Gain does, the final score of a feature should be the summation of the weighted parameters divided by the number of feature-value pairs. Assume there are

totally K features. For the kth feature with j_k feature-value pairs, the angles and p-values for the ith feature-value pair are a_{i1} and p_{i1} for the positive class, and a_{i2} and p_{i2} for the negative class, respectively. Then the score of the kth feature can be calculated through Equation (3) or (4).

$$\frac{\sum_1^{j_k}(w_1 \cos a_{i1} + w_2 \max((1-p_{i1}), p_{i2})}{j_k}. \quad (3)$$

$$\frac{\sum_1^{j_k}(w_1 \cos a_{i2} + w_2 \max((1-p_{i2}), p_{i1})}{j_k}. \quad (4)$$

If a feature-value pair is closer to the positive class, which means a_{i1} is less than 90 degrees, then Equation (3) is applied, where $\max((1-p_{i1}), p_{i2})$ would allow us to take the p-value with both classes into account. This is because that $(1-p_{i1})$ is the probability of the correlation between this feature-value pair and the positive class being true, and p_{i2} is the probability of its correlation with the negative class being false. Larger values of these two probabilities both indicate a higher level of reliability. On the other hand, if a_{i1} is larger than 90 degrees, which means the feature-value pair is closer to the negative class, then $\max((1-p_{i2}), p_{i1})$ will be used instead, that is Equation (4). w_1 and w_2 are the weights assigned to these two parameters. The pseudo code of integrating the angle value and p-value as a feature scoring metric (considering Jeffrey-Lindley paradox) is shown as follows.

CALCULATE SCORE

```
for k = 1 to K
    for i = 1 to j_k
        if cos a_i1 > 0
            sum_k += w_1 × cos a_i1
            if count_i1 > 0.01 AND count_i2 > 0.01
                sum_k += w_2 × max((1 − p_i1), p_i2)
        else if cos a_i1 < 0
            sum_k += w_1 × cos a_i2
            if count_i1 > 0.01 AND count_i2 > 0.01
                sum_k += w_2 × max((1 − p_i2), p_i1)
        else
            sum_k += 0
    end
    score_k = sum_k / j_k
end
```

Finally, after getting a score for each feature, a ranked list would be generated according to these scores, and then different stopping criteria can be adopted to generate a subset of features. The stopping criteria used will be discussed in Section IV since it depends on the design of each experiment. Afterwards, classifiers can be trained based on the selected features.

IV. Experiments and Results

Our proposed MCA-based feature scoring metric that integrates both correlation and reliability information is compared with four feature selection algorithms: information gain (IG), chi-square measure (CHI), correlation-based feature selection (CFS), and relief (REF). In order to find a good metric for feature selection that can improve classification accuracy, reduce computational complexity, and enhance semantic interpretability, the proposed framework is evaluated using the benchmark data from TRECVID 2009 video semantic concepts with totally 48 features, and each instance is a frame or shot from a video. Eight highly imbalanced concepts are chosen to show the effectiveness of different feature selection methods on improving classification accuracy since the performance of the classifiers decreases enormously with an imbalanced data set. The concept numbers, names, and the positive to negative (P/N) ratio are shown in Table I.

TABLE I
CONCEPTS TO BE EVALUATED

No.	concept name	P/N ratio
1	chair	0.074
2	infant	0.013
3	traffic-intersection	0.014
4	airplain-flying	0.027
5	person-playing-soccer	0.007
6	people-dancing	0.017
7	boat-ship	0.058
8	singing	0.074

Three-fold cross validation is first applied to the whole dataset of each concept, which randomly splits the data into three sets with an approximately equal number of data instances and an equal P/N ratio. Then each fold uses two of three sets as the training data set and the remaining one as the testing data set. The final result is the average of these three folds. As shown in Fig. 2, MCA only takes nominal features, but the extracted features from raw videos are all numeric. In order to get nominal features, discretization on the training data set needs to be conducted, and then the same intervals are used to discretize the testing data set. As can be seen, the discretization methods chosen would affect the final classification result. But to the best of our knowledge [16] [17] and also testing results, so far no particular discretization method is clearly superior to the others for our data. Thus, the discretization method applied here is the standard one embedded in WEKA, which is minimum description length (MDL) [18] [19]. Next, all the five feature selection algorithms are performed on the discretized training data set, which also reduce the effect of discretization on our comparison. Then different ranked lists would be generated based on different algorithms, except CFS which automatically produces the preferred subset. For different concepts, the weights in Equations (3) and (4) are different. Considering both cosine angle value and p-value lie between $[0, 1]$, according to our experiment data, five trials of different ratios between w_1 and w_2, which are 0:1, 1:1, 1:2, 2:1, 1:0, are considered in the experiments to ensure the computational complexity is acceptable.

After applying these five feature selection methods, for ranker methods IG, CHI, REF and the proposed MCA-based

algorithm, the generated training data set and the corresponding testing data set are data with the sorted features, while for non-ranker CFS, the generated data is a subset of the data with pruned features. Then these five sets of data, one for each feature selection method, are run under five classifiers, namely Decision Tree (DT), Rule based JRip (JRip), Native Bayes (NB), Adaptive Boosting (Ada), and K-Nearest Neighbor classifier where K is 3 (KNN). The stopping criterion used for the ranker methods is backward elimination which prunes the sorted features one by one backward after each time of classification. Each time, the precision, recall and F1-score of each classifier based on a particular subset of the features can be obtained. To conduct a complete search, 48 features require each classifier to repeat 47 times of classification on the sequentially decreased feature subspace produced by each ranker method. Based on the classification results, different feature subsets could be chosen for different comparison focuses. For example, for each feature selection method, the subset that results in the highest F1-score can be chosen as the best subset, or the chosen subset could be the one with a minimum number of features but still produces relatively high classification results.

In order to evaluate the proposed MCA-based algorithm in terms of the improved classification accuracy, the number of features reduced as well as the ability to capture the semantics in the videos, the experiments contain three parts. The first part is to compare the classification performance of these five classifiers when they are trained and tested using the subset generated by each feature selection method. Since CFS gives out the subset directly, in order to compare with it and not bias to any ranker method, for each concept, the same size of subspace as in CFS is chosen to evaluate each method. In Tables II, III, and IV, the average performance measures of five classifiers are shown for each concept. From these three tables, it can be observed that our proposed method outperforms the other four methods in precision, recall, and F1-score measures when the same size of feature subspace is used. On average (avg), it achieves 7% increase of the F1-score measure, and the standard deviation (std) across three folds is comparable to REF which is the best in these four methods. It can also be seen that the performance of IG and CHI are quite similar, and CFS is comparable to them given the size of the subspace is chosen based on it, while REF performs the worst.

The second part is to compare the reduced portion of the feature space under the same F1-score value. Due to the experimental result that CFS eliminates nearly two thirds of the features and the maximum F1-score value for each concept it reaches is far below the other methods, and meanwhile the maximum F1-score values of the other four methods are very close, we compare the feature space selected by each method except CFS when they reach the same maximum or close to maximum F1-score value. Figures 3 to 7 show the size of the subspace of our feature selection algorithm (denoted as MCA) compared to IG, CHI and REF. As can be seen from these figures, among all five classifiers, REF tends to choose the most features among four methods, especially under DT and JRip. On the other hand, our algorithm removes at least half of the total number of features, and the average size of our feature subset is about 20% smaller compared to the other three methods over all the classifiers.

Last, we look into the features chosen by each method in the first part. That is, to compare the F1-score under the same size of feature subspace. Since it is not easy to see intuitively whether each feature is useful or not for classification, concept 5 ("person-playing-soccer") is chosen as an example to show the semantic interpretability of each method. In the 48 used features, there are low-level visual and audio features extracted directly from raw video data, and then some of these low-level features are used to construct a set of middle-level features which are motivated by high-level semantics. For "person-playing-soccer" concept, based on the research study in [20],

TABLE II
AVERAGE PERFORMANCE OF MCA-BASED FEATURE SELECTION METHOD

No.	MCA		
	pre	rec	f1
1	0.63	0.31	0.40
2	0.57	0.15	0.23
3	0.83	0.24	0.37
4	0.49	0.11	0.17
5	0.72	0.54	0.58
6	0.52	0.24	0.32
7	0.56	0.24	0.33
8	0.53	0.22	0.30
avg	0.61	0.26	0.34
std	0.21	0.02	0.03

TABLE III
AVERAGE PERFORMANCE OF IG AND CHI FEATURE SELECTION METHODS

No.	IG			CHI		
	pre	rec	f1	pre	rec	f1
1	0.58	0.26	0.35	0.56	0.26	0.35
2	0.53	0.09	0.16	0.59	0.09	0.15
3	0.82	0.19	0.31	0.84	0.20	0.31
4	0.41	0.06	0.09	0.42	0.06	0.10
5	0.66	0.44	0.45	0.64	0.45	0.45
6	0.47	0.17	0.24	0.49	0.16	0.22
7	0.54	0.22	0.30	0.53	0.20	0.28
8	0.48	0.18	0.25	0.48	0.18	0.24
avg	0.56	0.20	0.27	0.57	0.20	0.26
std	0.10	0.05	0.08	0.19	0.04	0.07

TABLE IV
AVERAGE PERFORMANCE OF REF AND CFS FEATURE SELECTION METHODS

No.	REF			CFS		
	pre	rec	f1	pre	rec	f1
1	0.51	0.24	0.33	0.60	0.27	0.36
2	0.31	0.07	0.12	0.42	0.09	0.14
3	0.53	0.16	0.17	0.83	0.19	0.30
4	0.25	0.04	0.05	0.32	0.04	0.07
5	0.44	0.42	0.43	0.62	0.42	0.45
6	0.35	0.07	0.11	0.50	0.19	0.25
7	0.53	0.21	0.29	0.48	0.22	0.27
8	0.48	0.18	0.24	0.51	0.19	0.26
avg	0.44	0.18	0.21	0.53	0.20	0.26
std	0.30	0.02	0.04	0.25	0.07	0.11

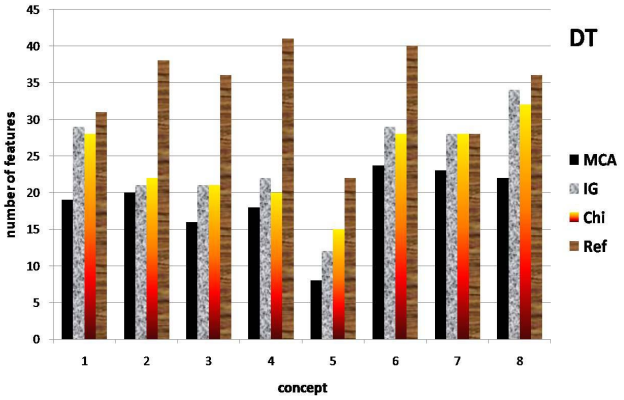

Fig. 3. Number of features used in Decision Tree

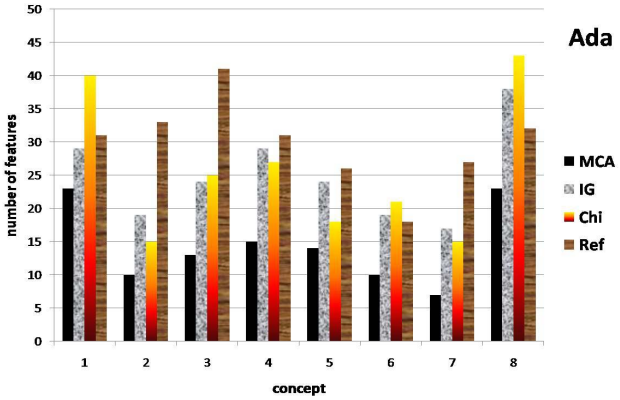

Fig. 6. Number of features used in Adaptive Boosting

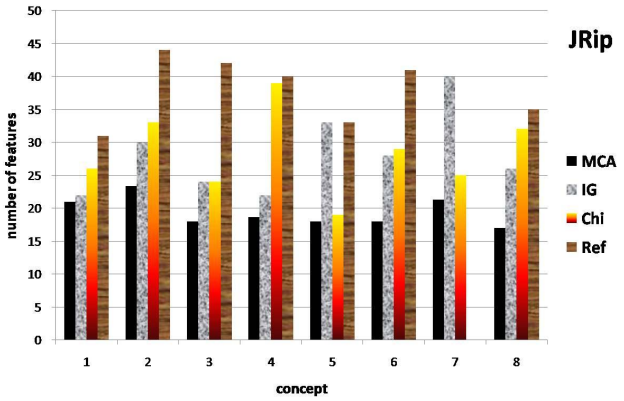

Fig. 4. Number of features used in JRip

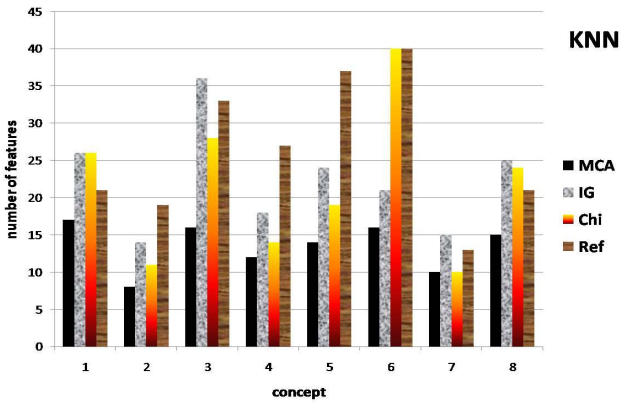

Fig. 7. Number of features used in K-Nearest Neighbor

a middle-level feature called grass-ratio defined as the ratio of grass area over the frame size is very useful for detecting the soccer videos. According to the ranked list generated by each method, grass-ratio is in the selected subset for our method since its corresponding score is high. However, none of the other methods include grass-ratio in their selected features. Moreover, 16 extracted audio features are also supposed to help differentiate soccer videos from some other concepts considering there usually is a speech along with the video shot. 4 volume related and 7 energy related features aim to distinguish speech from music. 4 Spectrum Flux features are often used in a quick classification of speech and non-speech, and the last one is zero crossing rate. Experiments show that IG, CHI and REF do not select any of the audio features, and CFS selects one that is the standard deviation of the Spectrum Flux. Our proposed method, on the other hand, selects four to six audio features in each fold. For example, the standard deviation of the volume for speech is much larger than it is for music because the shot pauses in our speech, resulting in many local minima which are close to zero interspersed between high values. Thus, by giving high ranking scores to features with more semantic information, from Tables II, III and IV, it can be seen that the F1-score of our method outperforms the other methods in the experiments by 13% for this concept. This shows that our proposed method prefers those features that are more semantically meaningful to humans and also more closely related to the semantics of the concepts. Thus it has more capability to map from low-level or middle-level features to high-level concepts.

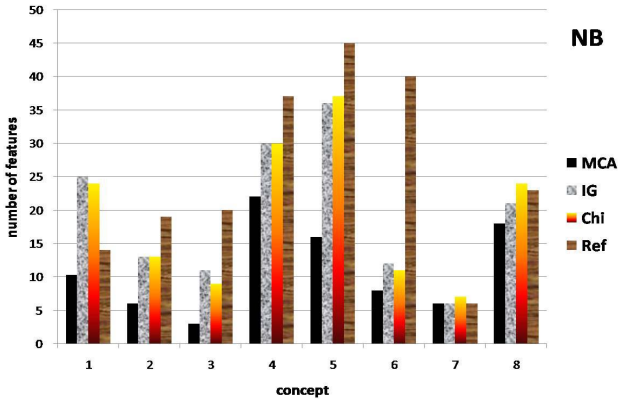

Fig. 5. Number of features used in Native Bayes

V. Conclusions

Feature selection has been considered as an important process in content-based multimedia retrieval to remove irrelevant or redundant features, which can improve model performance, reduce computational complexity, and enhance semantic interpretability. The angles from MCA has been well utilized as an indicator of correlation between features and classes, and also an indicator of the contribution of the features. In this paper, the geometrical representation of MCA is reused and p-value is taken as a measure of reliability of the relations between features and classes. An effective and integrated scoring metric for feature selection which takes both angle values and p-values into account is developed to score the features. A ranking list of features can be generated according to the scores and then a feature subset can be selected based on different criteria. In our experiment, eight highly imbalanced video concepts from TRECVID 2009 data are used to evaluate our proposed framework. The results show that compared to information gain, chi-square measure, correlation-based feature selection, and relief feature selection methods, the ranked list of features generated by our proposed framework ensures approximately 7% improvement of the F1-score values and 20% further reduction of the size of feature subspace over five popular classifiers, and meanwhile helps to bridge the semantic gap by giving high ranks to those features with more semantic information.

Acknowledgement

For Shu-Ching Chen, this work was supported in part by NSF HRD-0833093.

References

[1] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[2] H. Liu, J. Sun, L. Liu, and H. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol. 42, no. 7, pp. 1330 – 1339, 2009.

[3] J. Hua, W. D. Tembe, and E. R. Dougherty, "Performance of feature-selection methods in the classification of high-dimension data," *Pattern Recognition*, vol. 42, no. 3, pp. 409–424, 2009.

[4] Y. Lu, I. Cohen, X. S. Zhou, and Q. Tian, "Feature selection using principal feature analysis," in *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, 2007, pp. 301–304.

[5] M. J. Greenacre and J. Blasius, *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall/CRC, 2006.

[6] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Correlation-based video semantic concept detection using multiple correspondence analysis," in *Proceedings of the 10th IEEE International Symposium on Multimedia*, 2008, pp. 316–321.

[7] ——, "Effective feature space reduction with imbalanced data for semantic concept detection," in *SUTC '08: Proceedings of IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, 2008, pp. 262–269.

[8] J. Kludas, E. Bruno, and S. Marchand-Maillet, "Can feature information interaction help for information fusion in multimedia problems?" *Multimedia Tools and Applications*, vol. 42, no. 1, pp. 57–71, 2009.

[9] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed*. Morgan Kaufmann, 2005.

[10] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Information Processing and Management*, vol. 42, no. 1, pp. 155–165, 2006.

[11] D. C. Manning, P. Raghavan, and H. Schutze, "Text classification and naive bayes," in *Introduction to Information Retrieval*. Cambridge University Press, 2008, pp. 253–287.

[12] Y. G. Jiang, J. Yang, C. W. Ngo, and A. G. Hauptmann, "Representations of keypoint-based semantic concept detection: a comprehensive study," *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 42–53, 2010.

[13] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, pp. 359–366.

[14] Y. Sun, "Iterative relief for feature weighting: Algorithms, theories, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1035–1051, 2007.

[15] D. Lindley, "A statistical paradox," *Biometrika*, vol. 44 (1-2), pp. 187–192, 1957.

[16] L. A. Kurgan and K. J. Cios, "Caim discretization algorithm," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 2, pp. 145–153, 2004.

[17] C.-J. Tsai, C.-I. Lee, and W.-P. Yang, "A discretization algorithm based on class-attribute contingency coefficient," *Information Sciences*, vol. 178, no. 3, pp. 714–731, 2008.

[18] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1993, pp. 1022–1027.

[19] I. Kononenko, "On biases in estimating multi-valued attributes," in *Proceedings of the 14th international joint conference on Artificial intelligence*, 1995, pp. 1034–1040.

[20] M. Chen, S.-C. Chen, M.-L. Shyu, and K. Wickramaratna, "Semantic event detection via temporal analysis and multimodal data mining," *IEEE Signal Processing Magazine, Special Issue on Semantic Retrieval of Multimedia*, vol. 23, pp. 38–46, 2006.