SIViP (2008) 2:241–250
DOI 10.1007/s11760-007-0049-9

ORIGINAL PAPER

Feature selection for content-based image retrieval

Esin Guldogan · Moncef Gabbouj

Received: 18 October 2006 / Revised: 18 December 2007 / Accepted: 19 December 2007 / Published online: 19 February 2008

Abstract In this article, we propose a novel system for feature selection, which is one of the key problems in content-based image indexing and retrieval as well as various other research fields such as pattern classification and genomic data analysis. The proposed system aims at enhancing semantic image retrieval results, decreasing retrieval process complexity, and improving the overall system usability for end-users of multimedia search engines. Three feature selection criteria and a decision method construct the feature selection system. Two novel feature selection criteria based on inner-cluster and intercluster relations are proposed in the article. A majority voting-based method is adapted for efficient selection of features and feature combinations. The performance of the proposed criteria is assessed over a large image database and a number of features, and is compared against competing techniques from the literature. Experiments show that the proposed feature selection system improves semantic performance results in image retrieval systems.

Keywords Feature selection · Mutual information · Intercluster analysis · Inner-cluster analysis · Majority voting · Content-based indexing and retrieval

This work was supported by the Academy of Finland, Project No. 213,462 (Finnish Centre of Excellence Program 2006–2011) .

E. Guldogan (✉) · M. Gabbouj
Institute of Signal Processing, Tampere University of Technology,
P.O. Box 553, 33101 Tampere, Finland
e-mail: esin.guldogan@tut.fi

M. Gabbouj
e-mail: moncef.gabbouj@tut.fi

List of symbols

$p(x)$	Probability density functions
$p(x, y)$	Joint probability density function
$I(X; Y)$	Mutual information
$H(X)$	Shannon's entropy
S	Correlation measure for evaluating the discrimination power of feature
c	Number of classes
δ	Correlation between clusters
f_{xi}	ith item in the cluster x
μ_x	Mean of cluster x
σ_x	Standard deviation of cluster x
N_x	Cardinality of clusters x
e_1	Eigen vector corresponding to the largest eigen value of the covariance matrix
π	The best representative feature vector
x_N	Set of feature vectors
x_i	Feature vector corresponding to the ith item in the cluster
x_{ij}	jth element of the feature vector corresponding to the ith item of the cluster
M	Mean vector
μ_j	Elements of M, mean values
Δ	Distance between π and M
n	Number of elements in the vectors π and M
d	Euclidean distance between cluster members
S_{w1}, S_{w2}, S_{w3}	Compactness measurements
r	Covering radius, distance from the center to the farthest item in the cluster
Π	Probability
υ_{MI}	Normalized numerical results from mutual information criterion

υ_{ICR}	Normalized numerical results from inner-cluster relation criterion
υ_{PPMC}	Normalized numerical results from Pearson's product-moment correlation criterion
ν_{f_i}	Votes for each feature
F	Number of features in the FSRL list
α_i	Weights of the features in retrieval
R_i	Rank of the ith feature in FSRL list
ω_i	Weight of item i in SPFL list
ω_j	Weight of item j in FL list

1 Introduction

Content-based image indexing and retrieval (CBIR) systems often analyze image content via so-called low-level features, such as color, texture, and shape [1–4]. To achieve significantly higher semantic retrieval performance, recent systems tend to combine low-level features with high-level features that contain perceptual information for human. However, such combinations increase time and memory requirements together with retrieval complexity of feature extraction process.

Because of high memory and processing power requirements, CBIR has not been widely implemented on limited platforms, such as mobile devices or distributed systems. However, multimedia capabilities of all computing devices are growing steadily. Recently, multimedia became one of the key features of these devices for end-users. Hence, the necessity of multimedia services running on these platforms has arisen, where image indexing and retrieval is one of the most important challenges. Performance optimization of indexing and retrieval processes plays an important role for providing successful CBIR services for such limited systems. Feature selection is one of the key challenges for optimization of CBIR systems [5–8]. It refers to selecting the most important features and their combinations for describing and querying items in the database to reduce retrieval (time and computational) complexity while maintaining high retrieval performance. Moreover, it helps end-users by automatically associating proper features and weights for a given database.

Feature selection has been a popular research topic in pattern recognition. It has been applied to various research fields such as genomic data analysis, classification of network data, categorization of medical data, speech recognition, etc. [9–14]. However, assessments of feature performance and feature selection methods for CBIR have to be carried out in slightly different ways from classification and categorization. Decision errors are utilized for this purpose in classification. There is no unsupervised method to evaluate retrieval results of a CBIR system for assessing the semantic feature performance.

In this article, we mainly propose two criteria for feature evaluation and a method for feature selection that have not been addressed by earlier studies particularly in CBIR context:

- A new criterion based on categorized member relation within the same cluster from labeled training data to better understand the description power of the feature for each cluster.
- A new criterion based on the discrimination power of the features calculated using Pearson's product–moment correlation (PPMC) for defining correlations between different classes.

Using mutual information, intercluster and inner-cluster relations, a majority vote is applied on the results of these criteria as a decision mechanism to select appropriate features.

The organization of the article is as follows. Section 2 presents relevant feature selection methods in details. A majority voting method is described in Sect. 3. Experimental results are given in Sect. 4, and finally Sect. 5 provides concluding remarks and discussions.

2 Feature selection

Feature selection can be defined as selecting the combination of features among a given larger set that describes a particular data collection best. It has been a popular research topic since 1970's in pattern recognition and applied to several research fields.

In data mining, feature selection algorithms are divided into three categories: filters, wrappers, and hybrid methods. Filters use general characteristics of the data independently from the classifier for the evaluation process. The evaluation process is classifier-dependent in wrapper methods. Finally, hybrid models use both filtering and wrapping methods for improving the performance of the selection process.

Evaluating the discrimination power of the individual feature is a key operation in feature selection processes. Several methods may be used to evaluate the discrimination power of a feature, where mutual information is the most commonly used method [13–16]. Vasconcelos and Vasconcelos [8] used maximal divergence for feature selection in image retrieval; Ding and Peng [14] used mutual information for feature selection from Microarray gene expression data.

Intercluster relations are also used for medical image feature data evaluation in [5]. In this study, we utilize three criteria for different attributes of feature–data relations. Mutual information is used for measuring the feature and data relations. Intercluster and inner-cluster affinity characterizes the relationship between features and classes; thus, they are useful for evaluating the discrimination and description power of the feature, respectively.

The criteria, how and why they are used in our feature selection approaches are described in the following subsections.

2.1 Mutual information

Mutual information (MI) measures how much knowledge two variables carry about each other. It is the difference between the sum of the marginal entropies and their joint entropy. The mutual information of two independent items is always zero.

In [13], maximum dependency criterion based on mutual information is used for feature selection, and experimented with various data classification accuracies. Conditional mutual information is used for speech recognition in [15]. In this study, we use mutual information, where Shannon's entropy is utilized.

Definition Let x and y be two random variables, $p(x)$ and $p(y)$ be their probability density functions and $p(x, y)$ be their joint probability density function. Then their mutual information is defined as follows:

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{1}$$

The relationship between entropy and mutual information can be described as follows.

Let $H(X)$ denote Shannon's entropy of X, then

$$H(X) = -\int p(x) \log(p(x))dx \tag{2}$$

Then entropy is related to mutual information as follows:

$$I(X;Y) = H(X) - H(X|Y), \tag{3}$$

$$I(X;Y) = H(Y) - H(Y|X), \quad \text{or} \tag{4}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{5}$$

As a feature selection criterion, the best feature will maximize the mutual information $I(X;Y)$, where X is the feature vector and Y is the class indicator.

2.2 Pearson's product–moment correlation between data clusters

Intercluster information is widely utilized in cluster analysis for classifying data by using multiple features. The attributes of affinity between clusters (intercluster) represent the discrimination characteristics of a feature for a given data set. The data space is clustered with each feature individually, and the sums of correlations or distances are compared for evaluating the features. The criterion measuring class separability represents how the distances among the means of classes are maximized. Usually, this method is not used alone in feature selection approaches. The results become more reliable for discrimination if it is supplemented with inner-cluster information.

In [5], information concerning cluster compactness and cluster separability is used for unsupervised feature selection for content-based medical image retrieval. Class separability is also used in [17] for feature selection in a handwritten character recognition system. Class correlation is one of the criteria proposed in this study for evaluating the discrimination characteristics of a feature for a given set of data. It measures how cluster means are scattered with respect to each other. Large distances between clusters lead to better cluster discrimination. We use the following correlation measure for evaluating the discrimination power of each feature separately:

$$S = \sum_{x=1}^{c} \sum_{y=x+1}^{c} \delta(x,y) \tag{6}$$

where c represents the number of classes and δ represents the correlation between clusters x and y that are the numbers referring to cluster names or labels.

We use PPMC for defining the correlation between cluster means. The PPMC coefficient is the most commonly used measure of correlation in machine learning. It is calculated by summing up the products of the score deviations from the mean. We will use the following expression for the cluster correlation;

$$\delta(x,y) = \frac{\sum_{i=1}^{N_x} (f_{xi} - \mu_x)(f_{yi} - \mu_y)}{N_x N_y \sigma_x \sigma_y} \tag{7}$$

where f_{xi} and f_{yi} represents the ith item in the cluster x and y, μ_x and $\mu_{y,}\sigma_x$ and σ_y are the means and the standard deviations, N_x and N_y are the cardinality of clusters x and y, respectively. Clusters x and y are assumed to have equal number of elements to be compared by PPMC.

2.3 Inner-cluster relation based on first principal component

Inner-cluster scatter information is a widely used criterion in cluster analysis. The main objective of the inner-cluster analysis is to better understand the existing pattern in a given data space. It is often difficult to make assumptions about the cluster shape and distribution. Irregular shape clusters are particularly problematic. Inner-cluster information is also used for feature selection [5,18]. The most common inner-cluster information is "compactness," which is a measure of the similarity and closeness of the elements in a cluster. We use inner-cluster information as a criterion for feature selection in this study. If the elements of a cluster are close to each other in the represented feature space, or if the cluster is tight and compact, then the feature is considered as descriptive for the cluster.

We propose a new measure for inner-cluster information, inner-cluster relation (ICR), which represents the inner-cluster scatter information using the principal component information of the cluster. It is also related to the closeness of cluster elements similar to compactness.

ICR can be obtained by performing the following steps for a given set of feature vectors corresponding to a cluster of N elements:

Step 1 The aim of this step is to derive a feature vector (π) that represents the given cluster in terms of direction and characteristic by applying principal component analysis, which is a method for identifying patterns and highlighting the relations of the cluster elements. The first principal component e_1 is the eigen vector corresponding to the largest eigen value of the covariance matrix of the cluster. This process is principal component analysis and is applied to the feature vectors of the items in the cluster. The best representative feature vector π can be constructed using the following formula:

$$\pi = e_1 X \tag{8}$$

where X is a matrix containing set of feature vectors (x_N) for a cluster, and x_i represents the feature vector corresponding to the ith item in the cluster. x_{ij} is the jth element of the feature vector corresponding to the ith item of the cluster.

Step 2 In this step, we try to get information about the distribution of the cluster elements using the distance (Δ) between the representing feature vector π and mean vector M. The elements of M are the mean values μ_j of the feature vectors in the cluster calculated as follows:

$$\mathrm{M} = \left\{ \mu_j | \mu_j = \frac{1}{N} \sum_{i=0}^{N-1} x_{ij} \right\} \tag{9}$$

where N is the number of items in the cluster.

The sum of distances Δ is calculated with the following formula:

$$\Delta = \frac{1}{n} \sum_{i=0}^{n-1} (\pi_i - \mu_i)^2 \tag{10}$$

where n is the number of elements in the vectors π and M, and is also equal to feature vector dimension.

Step 3 In this step, the Δ value is normalized by the average distance between cluster elements to improve performance of the criterion on the clusters, where the cluster shape is not symmetric and the cluster distribution is not even. Finally, ICR is obtained as follows:

$$\mathrm{ICR} = \frac{\Delta}{\frac{2}{N(N-1)} \sum\limits_{i=0}^{N-2} \sum\limits_{j=i+1}^{N-1} d(x_{i,} x_j)} \tag{11}$$

where d is the Euclidean distance between cluster members and N is the number of items within the cluster.

2.3.1 Comparison of ICR and compactness

In this article, we rely on three different compactness definitions $S_{w1,} S_{w2}$, and S_{w3} used in the field of feature selection and cluster analysis, to compare the proposed method. The compactness criteria below approximately measures how scattered the cluster members are from their cluster means. The following equations present the definitions:

$$S_{w1} = \sum_{i=0}^{N-1} \|x_i - \mu\|^2, \tag{12}$$

where x_i is the ith member of the cluster and μ is the mean of the cluster [19].

$$S_{w2} = (\mu_d + \sigma_d + r), \tag{13}$$

where μ_d and σ_d are the mean and standard deviations of the Euclidean distances between the cluster members respectively. r is the covering radius, which is the distance from the center to the farthest item in the cluster [20].

$$S_{w3} = \mathrm{trace}(s_{w3}), \quad \text{where } s_{w3} = \Pi \Sigma \tag{14}$$

where Π is the probability and Σ is the covariance matrix of a cluster [5].

In Fig. 1, a sample data set consisting of subjectively similar images is represented by four different sets of two-dimensional features defined as Feature-1, -2, -3, and -4, and the clusters Cluster-1, -2, -3, and -4 are constructed, respectively, for each feature set. x and y axes of Fig. 1 represent the feature values and spatially closer cluster elements in the figure represent semantically related images. In the figure, features are energy and entropy values of Gray Level Co-Occurrence Matrix texture features [21]. In this example, we do not consider the semantic gap between low-level features and high-level objects. In this respect, construction of Cluster-1 can be considered more successful than that Cluster-2 and Cluster-4, since the elements in the Cluster-1 are spatially closer to each other than that in Cluster-2 and Cluster-4. Comparing the ICR values ($\mathrm{ICR}_{\text{Cluster-1}} < \mathrm{ICR}_{\text{Cluster-2}}$) leads to the same conclusion; while, the other three compactness factors depicts the opposite. Moreover, Cluster-3 elements can also be considered spatially closer than elements in Cluster-4. ICR values indicate the same ($\mathrm{ICR}_{\text{Cluster-3}} < \mathrm{ICR}_{\text{Cluster-4}}$); however, other three compactness criteria show that the Cluster-4 is the most compact cluster within four

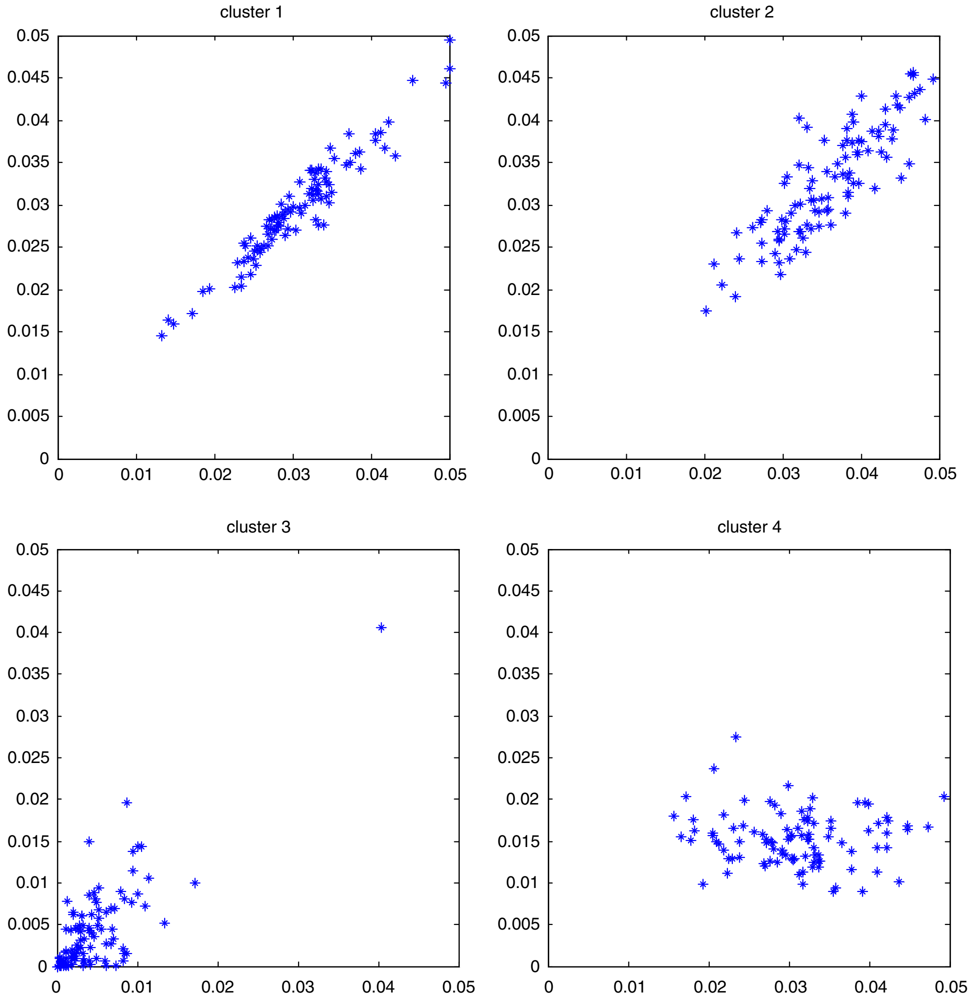

Fig. 1 A sample data set represented with four different sets of features

sample clusters. ICR gives better performance by the normalization of Δ value with the average distance between cluster elements to improve performance of the criterion on the clusters, where cluster shape is not symmetric and distribution is not even. It can be observed from Fig. 1 that ICR reveals better performance for describing the cluster distribution.

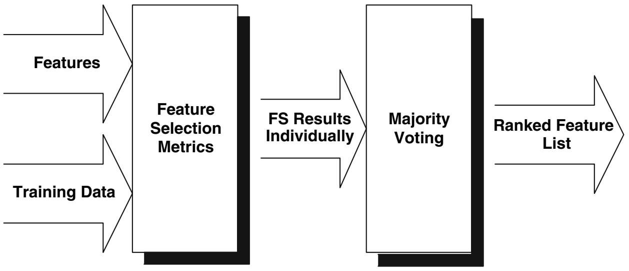

Fig. 2 Overview of the proposed feature selection system

3 Majority voting for features selection and weighting

Voting is a common classifier combination technique used in various disciplines, particularly in multi-classifier combination for pattern recognition [22,23]. In general, voting may be used as a black box and it does not require additional internal information for the decision implementation. Furthermore, it is a simple and effective method used in real-world applications as well as in social life as voting by majority.

In this study, majority voting is adopted in the feature selection process (Fig. 2). Majority voting selects the candidate having the largest amount of votes. We use voting method for sorting and selecting the features. Different from the categorization problem, the output of the decision-making black box is a list of features, which are sorted in descending order according to corresponding votes. The first feature in the output vote list represents the most important and powerful feature discriminating the associated data.

Feature voting for ranking and weighting works as follows:

- Voting gets the normalized numerical results from each individual criterion MI, ICR, and PPMC defined as υ_{MI}, υ_{ICR}, and υ_{PPMC} (see Sect. 4.2).

- Votes are calculated for each feature using the following formula

$$\nu_{f_i} = \sum(\upsilon_{\mathrm{MI}}(f_i) + \upsilon_{\mathrm{ICR}}(f_i) + \upsilon_{\mathrm{PPMC}}(f_i)) \tag{15}$$

- The features are sorted according to their ν_{f_i} values, and then the output feature selection ranking list (FSRL) is constructed.

3.1 Feature Selection by Voting

After obtaining the FSRL, an appropriate number of features are recommended to the end-user using the votes of the features in FSRL. Alternatively, appropriate number of features can be used internally by the system, without the user's knowledge, for unsupervised retrieval. In the ideal case, most definitive and discriminative features for querying a certain database are sorted and listed in the FSRL. The number of features that will be used for a query can be defined in a supervised way. On the other hand, the system gives a certain number of features to the end-user for an unsupervised way for retrieval. The minimum gradient of the sorted list of votes in FSRL corresponding to the sharpest decrease can be used to define the threshold, where the features with higher votes are recommended for retrieval. The defined threshold will be the number of features that are recommended by the system to the end-user.

3.2 Feature Weighting by Voting

FSRL is a sorted list of features that suitably represent the data. Users of the feature selection system may utilize all features in the FSRL instead of eliminating any of the features for the retrieval existing in the CBIR system. Appropriate feature combination should be given to the user in order to improve the semantic retrieval performance of the CBIR system. The orders of the features are given in FSRL list and an automatic weighting can be introduced to the users as follows:

Assume that the sum of the weights of the features is equal to 1.

$$\sum_{i=1}^{F} \alpha_i = 1, \quad \text{where } F \text{ is the number of features in the FSRL list.} \tag{16}$$

The weights of the features can be calculated as follows:

$$\alpha_i = \frac{(F - R_i) + 1}{\sum_{i=1}^{F} i} \tag{17}$$

where R_i represents the rank of the ith feature in FSRL list, and F is the number of features.

4 Experimental results

4.1 Data sets and features

Well-categorized Corel image data sets are widely used in the literature [24]. We used Corel real-world image databases for training and testing. For testing the results, a Corel database with 10,000 images are used. These images are preassigned to 100 semantic classes each containing 100 images by a group of human observers. Some examples of the classes are autumn, balloon, bird, dog, eagle, sunset, and tiger. Another Corel image database including 1,000 images categorized in 10 equal size classes is used for feature selection (training). In the first set of experiments, the following low-level color, shape, and texture features are used: YUV, RGB, and HSV color histograms with 128, 64, and 16 bins [25], Gabor Wavelet texture feature [26], Gray Level Co-Occurrence Matrix texture feature with parameters 12 and 6 [21], Canny Edge Histogram [27], and Dominant Color with three colors [28]. In the second set of experiments, same training and test databases are used and only color features YUV, RGB, and HSV color histograms with 32, 16, and 8 bins are utilized.

Image databases used for CBIR purposes tend to be large and lead to high complexity for feature selection. Thus, training data needs to be used. Computational and storage complexity will be decreased if feature and class probabilities and class relations are obtained from the training data. However, construction of the training data is another issue as it has direct impact on the precision accuracy of the feature selection method. It is a difficult task to construct training data to model general-purpose CBIR databases. Such databases contain random and irregular number of classes.

Defining the training data In this study, the training data is selected in a supervised way for evaluation and assessment of the methods. However, feature selection method is intended for any type of image data. Corel database contains 100 classes, where 10 of them (Corel 1,000 image database) are selected as training data. Feature subset selection is not employed since each feature is passed through the criteria individually.

4.2 Calculation of global criteria values υ_{MI}, υ_{ICR}, and υ_{PPMC}

Each criterion is applied separately for each individual feature to express clearly its effects of the characteristics in the data set for evaluation. Global criterion values are the inputs to the majority voting for final decision mechanism of the feature selection system.

Mutual information is calculated as shown in Eq. 1 for each feature using the training database, and the global mutual information value is the sum of these values. PPMC

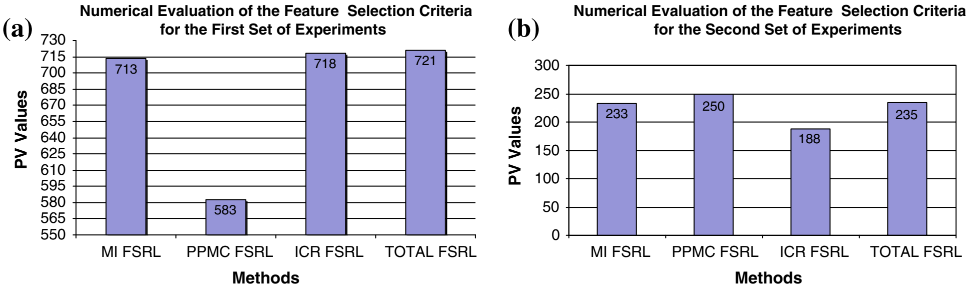

Fig. 3 **a**, **b** Numerical results of the proposed feature selection criteria

coefficients are calculated for every cluster combination for each feature, and sum yields the global PPMC value. ICR criterion is applied to each cluster in the data set with each feature individually. The sum of the values for each feature gives the global ICR value.

4.3 Assessment of the results

The objective of feature selection in CBIR context is decreasing complexity, improving usability, and particularly improving semantic performance. However, semantic performance of a CBIR system cannot be easily assessed. To evaluate the usefulness of the proposed feature selection method, we used a numerical method described in the following steps:

- Retrieval experiments are performed using each feature separately on the general database (so-called test database in this section) for obtaining average precision values. The test database has 100 classes each including 100 images. Thus, queries are performed with 500 images, five images from each class, and using each feature individually on the test database. Precision is defined as the ratio of the number of relevant records over the total number of retrieved records. It is usually expressed as a percentage. In these experiments, 36 retrieved images are taken into account for calculating the average precision.
- Semantic retrieval performances for corresponding features are recorded according to the precision values for verifying the feature selection method.
- Features are sorted according to the recorded retrieval performances. This step may also be expressed as sorting the features according to their representation level of the database. The sorted list is named as semantic performance feature list (SPFL). SPFL is used for evaluating the results of a feature selection system.
- The output of a feature selection system is a list of features so called feature list (FL) similar to the FSRL mentioned in Sect. 3. The proposed numerical assessment value referred to as performance value (PV) is calculated as follows:

$$\mathrm{PV} = \sum_{i}^{N} \sum_{j}^{N} \omega_i \omega_j \qquad (18)$$

where $\omega_i = N - \{\text{rank of item } i \text{ in SPFL}\} + 1$ represents the weight of item i in SPFL and $\omega_j = N - \{\text{rank of item } j \text{ in FL}\} + 1$, represents the weight of item j in FL.

Figure 3a, b represents the first and second set of experiments' PVs of MI (MI FSRL), PPMC (PPMC FSRL), as well as the proposed ICR (ICR FSRL) criteria separately, to compare it with the final FSRL results (TOTAL FSRL). It should be noted that the PV of the TOTAL FSRL is higher than other methods, which means that the final FSRL is closer to SPFL. In the best case, final FSRL should be equal to the SPFL. MI, PPMC, and ICR criteria represent different characteristics of the features on the data. Each of these criteria may work better than the others in different cases, as it can be seen from the first and second experiments. The semantic effects of these results on image retrieval are presented in the following section.

4.3.1 Comparisons of (compactness-ICR) and (separability-PPMC) with PV values

ICR and PPMC are also compared with S_{w1} compactness (see Sect. 2.3.1) [19] and class separability [5] using the proposed performance value. Figure 4 presents the retrieval results for each list of features constructed by the criteria. ICR outperforms compactness and PPMC outperforms class separability in terms of semantic retrieval performance based on the values in the figure.

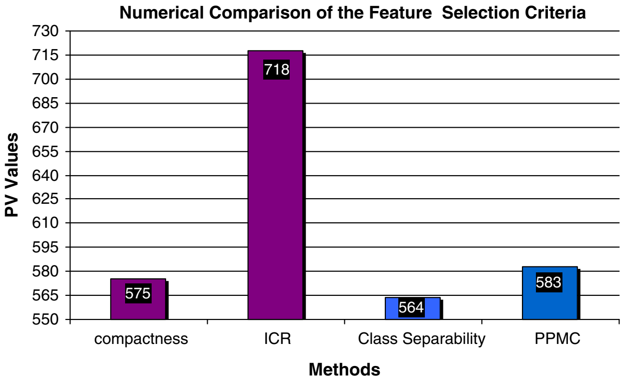

Fig. 4 Numerical comparisons of the feature selection criterions

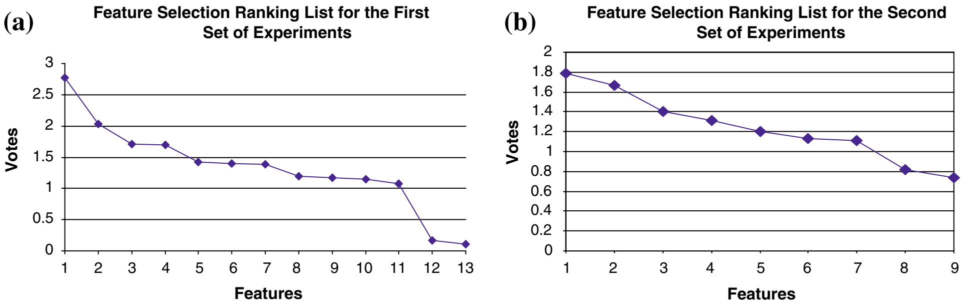

Fig. 5 **a**, **b** Votes of the features in final FSRL list

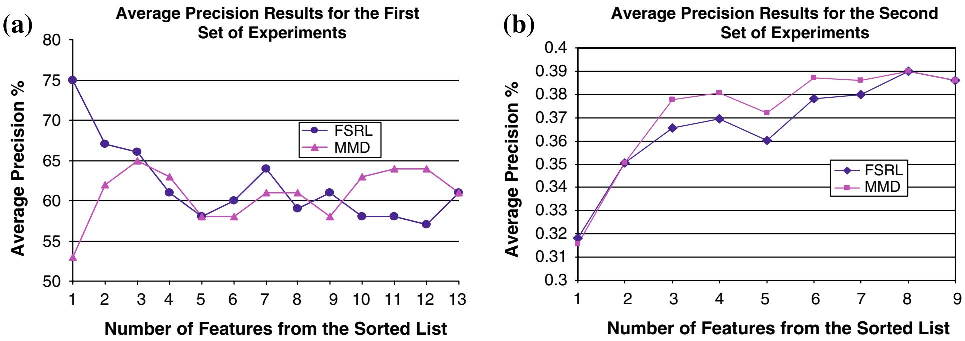

Fig. 6 **a**, **b** Average precision values of features employed in the experiments

4.4 Semantic retrieval results for image databases

In addition to the numeric results, average precisions are obtained from retrieval experiments for presenting semantic results. In the semantic retrieval performance experiments, Corel database including 100 classes, each including 100 images, is used. Five hundred queries are performed on the database by selecting five images randomly from each class. Average precision values are calculated according to these queries. The feature numbers given in Fig. 5 are taken from FSRL list. The features ranked higher in this list are considered as well-representative features for the database.

The proposed system is compared with the maximum marginal diversity (MMD) method for feature selection of image retrieval systems presented in [8]. MMD is used to construct a feature list, in which the features are ranked according to their representation success similarly to the construction of FSRL. Figure 5a, b illustrates the votes for each feature in the FSRL. In the first set of experiments shown in Fig. 5a, database has 13 features given in Sect. 4.1, and only the first feature is selected for image retrieval based on the votes, since the first feature (Feature-1) gives the maximum gradient. The recommended feature for this database is Gabor texture feature. In the second set of experiments shown in Fig. 5b, nine features are used, seven features are selected for image retrieval based on the votes by calculating the maximum gradient. The recommended features are sorted in FSRL list YUV color histogram (YUVCH) 16 bins and 32 bins, HSV color histogram (HSVCH) 32 bins, RGB color histogram (RGBCH) 32 bins, HSVCH 8 bins, YUVCH 8 bins, and HSVCH 16 bins according to representation power of the given database.

Figure 6a, b illustrates the results of retrieval using the features in the FSRL and MMD lists. The numbers in the x-axis of Fig. 5 refer to the rank of the underlying feature, e.g. 1 is for Feature-1, 2 is for Feature-2, etc. On the other hand, the same numbers in Fig. 6 refer to the number of features from the beginning of the list, e.g. 1 for the first feature, 2 for first two features each equally weighted, etc.

Ideally, a feature selection system should select the number of features, or x-axis value in other terms, which correspond to the highest average precision value in Fig. 6. The AP values should also tend to decrease or remain constant along the x-axis after the selected number of features. In the first example, the proposed system selects only one feature. It is verified by the results depicted in Fig. 6a as the precision starts to decrease after the first feature. The results with MMD in the same figure suggest that using three features gives the highest precision: HSVCH 128 bins, RGBCH 128, bins and HSVCH 64 bins. However, it is still lower than the precision obtained with the proposed feature selection system. In the second example, the proposed system selects seven features. It is verified by the results depicted in Fig. 6b as the precision slightly increases with the eighth feature and decreases after that. In Fig. 6b, MMD suggests five features: HSVCH 32 bins, YUVCH 32 bins, RGBCH 32 bins, HSVCH 16 bins, and RGBCH 8 bins. Final MMD average precision is lower

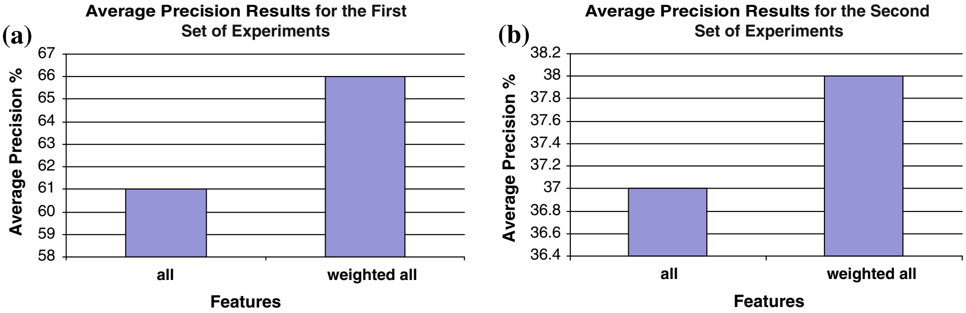

Fig. 7 **a**, **b** Average precision values of all features with different weights

than the result of the proposed system, although there are higher precisions shown partially in the MMD results.

An alternative way of using FSRL is an automatic determination of feature weights for retrieval. Feature weights are calculated automatically by the system using FSRL. In the experiments of Fig. 7a, b, first all existing features in the system are combined with equal weights, then the features with the proposed weights by the system are utilized. As shown in Fig. 7a, the average precision of 61% is increased to 66% when the proposed feature weighting is employed, and in Fig. 7b it is increased from 37 to 38%.

5 Discussions

The proposed feature selection and weighting method can be used to enhance semantic content-based image retrieval performance, to decrease retrieval process complexity, and to improve system usability for end-users. Our method uses three different criteria and a majority voting approach for the final decision-making step, where each criterion represents different feature characteristics.

The novelties of the proposed feature selection approach in a CBIR system are as follows:

- A new criterion based on categorized member relation within the same cluster from the labeled training data.
- A new criterion for defining the correlations of interclusters, which is based on PPMC and used for defining the discrimination power of the feature.
- The three different criteria with majority voting approach as a decision-making mechanism.

MI criterion refers to the amount of information a feature carries about the data. ICR criterion gives the feature description power for each labeled cluster, where members of a cluster are supposed to be similar and close to each other in the feature space. ICR is slightly similar to the compactness of clusters; however, its analysis accuracy is higher for irregular shape clusters. PPMC criterion represents the correlation between clusters, where each of them is uncorrelated in the ideal case. Instead of cluster separability, we use PPMC, since it discriminates the clusters better. Once the criteria are applied, each feature has normalized values, which will be the inputs to the voting system. Majority voting is adopted for the use of feature listing in CBIR context. It generates a sorted vote list referred to as FSRL with associated feature names. FSRL may be utilized to obtain one of the two different outputs to the end-users for the following use-cases:

- Recommended set of features will be used by the system automatically.
- Recommended weights for each feature will be given to the end-user or the system will use them automatically.

The choice of the use-case is directly related with the complexity. First use-case decreases the retrieval complexity, whereas the latter one improves the semantic performance without altering complexity, since all features are used in the retrieval process. Especially the first use-case may be applied on limited platforms having low capacity for computing features and all the features may not be used for CBIR.

The proposed system in CBIR can be used automatically or manually by the end-user. In unsupervised case, the system internally uses feature combinations and weights automatically. In supervised case, user selects the features and weights for retrieval.

Defining a suitable training data is a major challenge in this study. The proposed method should be performed on a representative training data for successful results in large image databases.

Flexibility and efficiency of the proposed approach allows it to be applied in various platforms and for types of data. The success of the proposed approach is verified with the experimental results on image databases.

6 Conclusions and future work

In this article, we explore the use of feature selection within a CBIR context. Two novel feature selection criteria based on inner-cluster and intercluster relations are proposed, and an efficient majority voting-based method is implemented for the selection and combination of features. The proposed method includes three main criteria for feature–data relation. These criteria produce results for each feature that are fed to majority voting as input. Each criterion is based on different associations of feature–data affinity to define the best discriminative and representative feature of the data. Two proposed criteria are compared with other state-of-the-art criteria through dedicated experiments, which show that the proposed methods improve retrieval performance. The proposed feature selection system is implemented as a black-box approach that gives flexibility for using it in different platforms. It may be performed on several types of databases and sets of features.

Moreover, assessment studies on the criteria will be carried out using different databases on different platforms in the future. In addition, this work may be extended to multimodal features for multimedia databases. Selection of the training data is a challenge in this study. Nonrepresentative sample training data hinders the generalization of the method. How to select appropriate training data is still an open problem to be studied in the future.

References

1. MUVIS: A system for content-based multimedia indexing and retrieval in multimedia databases. http://muvis.cs.tut.fi/
2. Pentland, A., Picard, R.W., Sclaroff, S.: Photobook: content-based manipulation of image databases. Int. J. Comput. Vis. **18**(3), 233–254 (1996)
3. Niblack, W., Barber, R., et al.: The QBIC project: querying images by content using color, textures and shape. In: Proceedings of SPIE Storage and Retrieval for Image and Video Databases, 1996, pp. 124–128 (1996)
4. Smith, J.R., Chang, S.-F.:VisualSEEk: a fully automated content-based image query system. In: Proceedings of ACM Multimedia, Boston, November 1996, pp. 87–98 (1996)
5. Dy, J.G., Brodley, C.E., Kak, A.C., Broderick, L.S., Aisen, A.M.: Unsupervised feature selection applied to content-based retrieval of lung images. IEEE Trans. Pattern Anal. Mach. Intell. **25**(3), 373–378 (2003)
6. Collins, R.T., Yanxi, L., Leordeanu, M.: Online selection of discriminative tracking features. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1631–1643 (2005)
7. Wei, J., Guihua, E., Qionghai, D., Jinwei, G.: Similarity-based online feature selection in content-based image retrieval. IEEE Trans. Image Process. **15**(3), 702–712 (2006)
8. Vasconcelos, N., Vasconcelos, M.:Scalable discriminant feature selection for image retrieval and recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–775 (2004)
9. Xing, E.P., Jordan, M.I., Karp, R.M.:Feature selection for high-dimensional genomic microarray data. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 601–608 (2001)
10. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowledge Data Eng. **17**(4), 491–502 (2005)
11. Jain, A., Zongker, D.: Feature selection: evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Mach. Intell. **19**(2), 153–158 (1997)
12. Koller, D., Sahami, M.:Toward optimal feature selection. In: Proceedings of the 13th International Conference on Machine Learning, pp. 284–292 (1996)
13. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–238 (2005)
14. Ding, C., Peng, H.: minimum redundancy feature selection from microarray gene expression data. In: Proceedings of the IEEE Computer Society Conference on Bioinformatics, 11–14 August 2003, pp. 523–528 (2003)
15. Ellis, D.P.W., Bilmes, J.A.: Using mutual information to design feature combinations. In: Proceedings of International Conference on Spoken Language Processing, ICSLP-2000, Vol. 3, Beijing, October 2000, pp. 79–82 (2000)
16. Hariri, S., Yousif, M., QuA, G.: New dependency and correlation analysis for features. IEEE Trans. Knowl. Data Eng. **17**(9), 199–1207 (2005)
17. Shi, D., Shu, W., Liu, H.: Feature selection for handwritten chinese character recognition based on genetic algorithms. IEEE Int. Conf. Syst. Man Cybernet. **5**, 4201–4206 (1998)
18. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Anal. Mach. Intell. **24**(3), 301–312 (2002)
19. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2001)
20. Kiranyaz, S., Gabbouj, M.: Hierarchical cellular tree: an efficient indexing scheme for content-based retrieval on multimedia databases. IEEE Trans. Multimed. **9**(1), 102–119 (2007)
21. Partio, M., Cramariuc, B., Gabbouj, M., Visa, A.: Rock texture retrieval using gray level co-occurrence matrix. In: Proceedings of 5th Nordic Signal Processing Symposium, October 2002. (2002)
22. Xu, L., Krzyzak, A., Suen, C.Y.: Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans. Syst. Man Cybernet. **22**(3), 418–435 (1992)
23. Lin, X., Yacoub, S.M., Burns, J., Simske, S.J.: Performance analysis of pattern classifier combination by plurality voting. Pattern Recognit. Lett. **24**(12), 1959–1969 (2003)
24. Corel Stock Photo Library, Corel, Ontario
25. Swain, M.J., Ballard, D.H: Color indexing. Int. J. Comput. Vis. **7**(1), 11–32 (1991)
26. Ma, W.Y., Manjunath, B.: Texture features for browsing and retrieval of image data. IEEE Trans. Pattern Anal. Mach. Intell. **18**, 837–842 (1996)
27. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **8**(6), 679–698 (1986)
28. Manjunath B., S., Ohm, J.-R., Vasudevan, V.V., Yamada, A.: Color and texture descriptors. IEEE Trans. Circuits Syst. Video Technol. **11**(6), 703–715 (2001)