Journal of Korea Multimedia Society Vol. 14, No. 12, December 2011(pp. 1544-1548)

Comparison of Feature Selection Processes for Image Retrieval Applications

YoungMee Choi†, MoonWon Choo††

ABSTRACT

A process of choosing a subset of original features, so called feature selection, is considered as a crucial preprocessing step to image processing applications. There are already large pools of techniques developed for machine learning and data mining fields. In this paper, basically two methods, non-feature selection and feature selection, are investigated to compare their predictive effectiveness of classification. Color co-occurrence feature is used for defining image features. Standard Sequential Forward Selection algorithm are used for feature selection to identify relevant features and redundancy among relevant features. Four color spaces, RGB, YCbCr, HSV, and Gaussian space are considered for computing color co-occurrence features. Gray-level image feature is also considered for the performance comparison reasons. The experimental results are presented.

Key words: Feature Selection, Image Retrieval, Sequential Forward Selection

1. INTRODUCTION

Feature selection is the process of identifying and removing as much of the irrelevant and redundant information as possible. The recent increase of dimensionality of data causes the deep efforts for selecting features with respect to efficiency and effectiveness[1]. It is a process of choosing a subset of original features so that the feature space is optimally reduced according to a certain evaluation criteria[2]. This concept could be applied especially to a variety of location-based game applications for mobile devices, which usually have limited resources. Data reduction improve the classification performance, the approximation function, and pattern recognition systems in terms of speed, accuracy and simplicity[3].

Feature selection process requires a search strategy to select candidate subsets and an object function to evaluate these candidates. Two broad categories of this process are referred to as wrapper and filter. The wrapper employs a statistical re-sampling technique using the actual target learning algorithm to estimate the accuracy of feature subsets. The filter operates independently of a learning algorithm, typically make use of all the available training data when selecting a subset of features. Since any predefined learning algorithm is not proposed, the filter model is considered more often than any other model because of its computational efficiency and good performance effectiveness without learning models[4]. However, these two approaches have many trade-offs in terms of speed, accuracy, availability of required data for modeling and simplicity. In this paper, the simple wrapper method is used to select the subset of features. The subset of features obtained through Sequential Forward Selection algorithm with k-nn classifier are compared with non-selection features. The training and testing image features are gen-

※ Corresponding Author : MoonWon Choo, 400-10 Manan-Gu, Anyang 8-dong, Anyang City, Kyunggi-Do, TEL : +82-31-467-8178, FAX : +82-31-449-0529, E-mail : mchoo@sungkyul.edu
Receipt date : Oct. 31, 2011, Revision date : Dec. 21, 2011
Approval date : Dec. 30, 2011
† Div. of Multimedia, Sungkyul University, Korea (E-mail: mchoo@sungkyul.edu)
†† Div. of Multimedia, Sungkyul University, Korea (E-mail: mchoo@sungkyul.edu)

erated by computing color co-occurrence features of images generated using several color spaces.

The outline of this paper is as follows. In section 2, the color co-occurrence features are described. Section 3 summarizes Sequential Feature Selection process and treats the symmetric uncertainty as the performance criteria for computing the degree of association. Section 4 shows the result of experiment. Section 5 briefly describes the conclusion and future research.

2. COLOR CO-OCCURRENCE FEATURES

The co-occurrence matrix is a well-known statistical tool for extracting second-order texture information from images and widely used for image retrieval applications [5-7]. This matrix can be thought of as an estimate of the joint *pdf* of gray level pairs in an image. Suppose the image I to be analyzed is an $M \times N$ dimensional matrix. An occurrence of some gray level intensity may be described by a matrix of relative frequencies $O_{\theta,d}(a,b)$, describing how frequently two pixels with gray levels a, b appear in the matrix separated by a distance d in direction θ. Non-normalized frequencies of co-occurrence as functions of angle 0° and distance can be represented formally as:

$$O_{0^{\circ},d}(a,b) = |\{((k,l),(m,n)) \in V : k-m=0, |l-n| = d, D(k,l)=a, D(m,n)=b\}| \quad (1)$$

where $|\{\cdots\}|$ refers to set cardinality, D is the target image block and $V = (M \times N) \times (M \times N)$. The gray-level parameters a, b are determined by the predefined image gray-levels. The distance metric ρ in these equations can be defined by $\rho[(k,l),(m,n)] = \max\{|k-m|, |l-n|\}$. This method can be extended to color images. Color images are usually coded in three channels.

The cross co-occurrences can be counted to utilized the correlation between two channels. An image in RGB color space may generate (R,R), (G,G), (B,B), (R,G), (R,B), (G,B) co-occurrence matrices. The dimension of each O is determined by the range of pixel values.

3. SEQUENTIAL FORWARD SELECTION (SFS)

The SFS algorithm can be simply formulated as follows (more details referred in [8]).

1. Starting from the empty set $\boldsymbol{Y_k} = \{\phi\}$;
2. Select the next optimal features $x^{*} = \arg\max_{x \notin Y_k}[J(Y_k + x)]$;
3. Update $Y_{k+1} = Y_k + x^{*}; k = k+1$;
4. Goto 2

That is, starting from the empty set, sequentially add the feature vector x^{*} that results in the highest objective function $J(Y_k + x^{*})$ when combined with the feature vectors Y_k that have already been selected. The object function J could be any proper function derived from optimization theory. This process performs best when the optimal subset has a small number of features.

The objective functions of filter approaches evaluates feature subsets by their information content, typically interclass distance, statistical dependence or information-theoretic measures. Those of wrappers are pattern classifier, evaluating feature subsets by their predictive accuracy by statistical re-sampling or cross-validation. In this paper standard SFS is used with k-nn classifier.

4. SYMMETRIC UNCERTAINTY(SU)

In general, a feature is good if it is relevant to the class but is not redundant to any of the other relevant features[9]. This relevance can be measured using correlation concept between two variables. There may be many approaches to compute the degree of correlation. One of them is the entropy, a measure of the uncertainty of a random

variable. The entropy of a variable X is defined as

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \quad (2)$$

and the entropy of X after observing values of another variable Y is defined as

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j)) \quad (3)$$

where $P(x_i)$ is the prior probabilities for all values of X, and $P(x_i|y_j)$is the posterior probabilities of X given the values of Y. The amount of which the entropy of X decreases reflects additional information about X provided by Y and is called information gain, given by

$$IG(X|Y) = H(X) - H(X|Y) \quad (4)$$

This symmetrical measure shows the relative degrees of correlations between pairs of variables. However, this measure is biased in favor of features with more values. To remedy this issue, symmetrical uncertainty is used, defined as follows[2],

$$SU(X,Y) = 2\left[\frac{IG(X|Y)}{H(X)-H(Y)}\right]. \quad (5)$$

An SU value of 1 indicates that using one feature other feature's value can be totally predicted and value 0 indicates two features are totally independent. The SU values are symmetric for both values. This symmetric uncertainty measures are usually used for objective functions of the feature selection algorithms specified as filter categories. In this paper it is adopted as the criteria of performance measures for feature selections.

5. EXPERIMENTS

Firstly the input color images for training and one class image are computed to get color co-occurrence matrices for each image, which consist of data set. Then SFC algorithm is applied to data set to get the optimal subset of feature vectors. Training image set used here consists of 100 images of indoor scenes with few specularities and taken under normal lightening conditions(see Fig. 1). They are two parts of the image sets, which were captured under different circumstances.

First set composes of images taken under indoor lighting only. The images of second set are to be tested were taken under indoor lightening mixed with outdoor lightening conditions.

To study of the validation of the suggested approach, several color models are investigated. RGB color space is standard color model for monitor without luminance information. YCbCr color space are mainly used in video settings and has three channels, such as Y(luminance), Cb(chrominance of blue), Cr(chrominanance of red). HSV(hue, saturation, value) color space corresponds better to how people experience color than RGB does. Additionally, the Gaussian color model(GCM), so called Kubelka-Munk theory model, is also considered [10,11,12]. For the purpose of performance

Fig. 1. Part of two sets of images for training and testing.

comparison, the gray-level feature is also included. The size of color images captured by SONY Cybershot DSC-T3 is 1944 x 2592, resized with 10:1 resolution ratio. The intensity levels of each color plane are adjusted to the range 0-19 for computational convenience. The distance parameter of co-occurrence feature is set to 1 and 8 directions are considered.

For comparison of performance, only the mean values of the SU of features from training images and test images are considered. The inter-feature SU values are computed for both kinds of images and the mean values are taken. The inter-class SU values are derived from training images by taking testing images as members of a different class.

For inter-class discrimination, it is assumed that the higher SU values to be considered as the higher degree of prediction or association of features with specific classes. Fig. 2 shows that after applying feature selection algorithm, the SU values are increased at the features from the most of the color spaces except HSV color space. RGB model apparently has the most useful power to separate two different classes, followed by Gaussian and gray-level model.

For inter-feature discrimination, it is assumed that the lower SU values to be considered as the higher degree of independence or lower redundancy of features belonging to specific class.

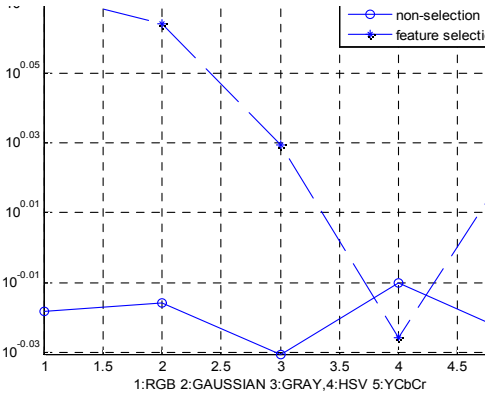


Fig. 2. The log-scaled mean of SU for inter-class discriminant.

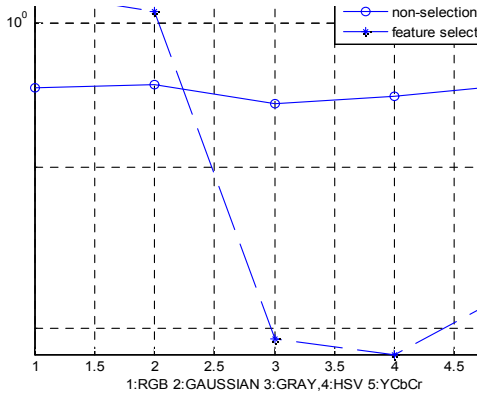


Fig. 3. The log-scaled mean of SU for inter-feature discriminant.

Fig. 3 shows that after applying feature selection algorithm the SU values are decreased at the features from gray-level, HSV, and YCbCr. However, RGB and Gaussian models lower the degree of redundancy, which may not be desirable for further classification processes.

6. CONCLUSION

Feature selection methods have been shown to be effective in removing redundant and irrelevant features, which improves learning algorithm's prediction performance and reduces the effects of curse of high dimensionality of data. However, this paper suggests that the color images data may be transformed into other color space for better classification or learning processes. But presupposed co-occurrence feature requires time-consuming computation and may not be suitable as image features for generic outdoor images. The result of this research is not complete enough to apply in general classification settings and may give some insights for easy preprocessing methods before applying second step for feature selection algorithms

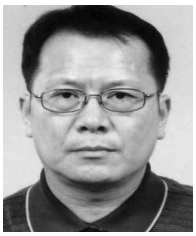
REFERENCES

[1] Sami B. and Jorma L., "Statistical Shape

Features in Content-Based Image Retrieval," *Proc. of ICPR2000*, Spain, September, 2000.
[2] Lei Yu and Huan Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," *Proceedings of ICML-2003*, Washington DC, 2003.
[3] Mark A., "Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper," *AAAI*, 1998.
[4] Peng H.C, " Feature selection based on mutual information: criteria of max_dependency, max-relevance, and min-redundancy," *IEEE Trans. on PAMI*, Vol.27, No.8, pp. 1226-1238, 2005.
[5] B. Chanda, B.B. Chaudhuri, and D. Dutta Majumder, "On image enhancement and thread selection using the gray-level co-occurrence matrix," *Pattern Recognition Lett.*, Vol.3, No.4, pp. 243-251, 1985.
[6] Gonzalez Woods, "Digital Image Processing," *Addison Wesley,* 1992.
[7] Haralick Shapiro, "Computer and Robot Vision Vol. 1," *Addison Wesley,* 1992.
[8] Pavel Krizek, "Feature Selection: Stability, Algorithms, and Evaluation," *Doctoral thesis,* University of Surrey, June 244, 2008.
[9] Augusto Destrero, "Feature selection for high-dimensional data," *Springer-Verlag,* 2008.
[10] Jan-Mak G., "Color Invariance," *IEEE Trans. On PAMI*, Vol.23, No.12, 2001.
[11] Kristen Hoffman, "Applications of the Kubelka-Munk Color Model to Xerographic Images," www.cis.rit.edu/research/thesis, 1998.
[12] SeokMin Chae, SungHak Lee and Myoung-Hwa Lee and Kyu-Ik Sohng, "Neutral point model of HVS for the Illuminant-adaptive White Balance Control of Displays," *Journal of Korea Multimedia Society*, Vol.13, No.5, pp. 674-683, 2010.

YoungMee Choi

She received her B.S. in Mathematics from Ewha Womans University and her M.S. in Numerical Analysis from Ewha Womans University, in 1979 and 1981, respectively. She received her Ph.D. in Computer Engineering in 1993 from Ajou University. She joined the Division of Multimedia Engineering at Sungkyul University as a professor. Her current research interests include the intelligent tutoring system, game AI and serious game.



MoonWon Joo

He received his B.S. in Mathematics from San Jose State University and his M.S. in Computer Science from New York Institute of Technology, in 1985 and 1987, respectively. He received his Ph.D. in Computer Science in 1996 from Stevens Institute of Technology. He joined the Division of Multimedia Engineering at Sungkyul University as a professor. His current research interests include the image processing and serious game.