Multi-module Image Classification System*

Wonil Kim[1], Sangyoon Oh[2,**], Sanggil Kang[3], and Dongkyun Kim[1]

[1] College of Electronics and Information Engineering at Sejong University, Seoul, Korea
wikim@sejong.ac.kr, kdk0909@paran.com
[2] Computer Science Department at Indiana University, Bloomington, IN, U.S.A.
ohsangy@cs.indiana.edu
[3] Department of Computer Science, The University of Suwon, Gyeonggi-do, Korea
sgkang@suwon.ac.kr

Abstract. In this paper, we propose an image classification system employing multiple modules. The proposed system hierarchically categorizes given sports images into one of the predefined sports classes, eight in this experiment. The image first categorized into one of the two classes in the global module. The corresponding local module is selected accordingly, and then used in the local classification step. By employing multiple modules, the system can specialize each local module properly for the given class feature. The simulation results show that the proposed system successfully classifies images with the correct rate of over 70%.

1 Introduction

The fast developing digital multimedia technology enables us to access much more information through the Internet and TV than any time before. The information is easily accessible, but it brings another side: how we store and manage these rich digital videos and images so that we can collect the proper information whenever and wherever we want. Because of recent hundreds of TV sports news programs broadcasted from all over the world, TV viewers need to choose the most interesting news and watch its highlight channel. Also, in order to manage a digital library for sports videos and images, we need an automatic image classification system.

The main purpose of this paper this paper is applying MPEG-7 to sports images for feature extraction and classification system. By analyzing MPEG-7 descriptors, we create a prototype system that can be used for sports image classification techniques under visual environments, and introduce an effective methodology of image classification via experiments. Our approach we present in this paper employs multi-module neural networks for the sports image classification system. The input value for the neural network is one of the values of visual features extracted by MPEG-7 descriptors.

In the next section, we discuss several related methods of the image classification including neural network and feature extraction using MPEG-7 descriptors. Then, we propose a Multi-Module Image Classification System in Section 3 and then discuss some simulation environments and results in Section 4. We conclude in Section 5.

* This paper is supported by Seoul R&BD Program.
** Author for correspondence : +1-812-856-0751.

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 498 – 506, 2006.

2 Related Works

2.1 Multi-module Approach

Given a complicated input-output mapping problem, it is hard to design a single network structure and solve the problem with this architecture. It usually takes a long time to train a monolithic network and may not produce a good generalization result. Since the mapping may be realized by a local method which captures the underlying local structure [1], there has been considerable research [2, 3, 4, 5, and 6] designed to take advantage of a modular network architecture. In this paradigm, each module is assigned to a specific (local) area and focuses only on its special area. Learning is more efficient when a neural network is organized in this way.

A neural network is said to be modular if the computation performed by the network can be decomposed into two or more modules that operate on distinct inputs without communicating with each other [7]. The outputs of the module are mediated by an integrating unit that may not feed information back to the modules. In particular, the integrating unit decides how the outputs of the modules should be combined to form the final output of the system, and decides which modules should learn which training patterns. Modular networks utilize the principle of divide and conquer, which permits us to solve a complex computational task by dividing it into simpler subtasks and then combining their individual solutions [8].

The use of a local method offers the advantage of fast learning and therefore requires relatively few training iterations to learn the task. Alternatively, an approximation may be realized using a global method that captures the underlying global structure of mapping. The use of global methods offers the potential advantages of smaller storage requirement and better generalization performance. The use of a modular approach may also be justified on neurobiological grounds. Modularity appears to be an important principle in the architecture of vertebrate nervous systems, and there is much that can be gained from the study of learning in modular networks in different parts of the nervous system.

2.2 Image Classification

Image classification is the core process of digital image analysis. It is used in many areas like remote sensing and image retrieval. Remote sensing is the acquisition of meaningful information from an object by a recording device that is not in physical or intimate contact with the object. For example, image classification is applied to a data interpretation process of remotely acquired digital image by a Geographic Information System (GIS).

The image retrieval also uses image classification. A user requests an image by query and it returns an image (or a set of ordered images) from its image database by matching features, like color histogram and textual measures, of a query image with those of the database images. Image classification is also used to create image databases and adding images to it for the image retrieval system. The system extracts semantic description from images and putting them into semantically meaningful categories. We focus our related work survey on this kind of systems.

For content-based image retrieval systems, one of the classic classification problems is city images vs. landscapes. Gorkani and Picard [9] separate urban images and

rural images using a multiscale steerable pyramid to find dominant orientations in four by four subblocks of image. They classify the image as a city scene if enough subblocks have vertical orientation tendency.

Vailaya et al. [10] also use city vs. landscape images to show how the high level classification problem (urban images vs. rural images and indoor vs. outdoor images) can be solved from low-level features geared toward the particular classes. They have developed a procedure to measure qualitatively the saliency of a low-level visual feature towards a classification problem based on the plot of the intra-class and inter-class distance distributions. They determine that the edge direction-based features have the most discriminative power for the classification problem.

Indoor-outdoor problem studied by Szummer and Picard is a variant scene classification problem. Their paper [11] shows the performance improvement by computing features on subblocks, classifying these subblocks, and combining results in stacks. Features in the study include histograms in the Ohta color space [1], multiresolution, simultaneous autoregressive model parameters, and coefficients of a shift-invariant DCT. For combining the results of subblocks, they compare the classification performance of the usage of a simple classifier, k-nearest neighbors, and that of other sophisticated classifiers, like neural networks and mixture of expert classifiers. In their test set results, the simple nearest neighbor classifier performs better.

Texture like a texture orientation used in Ref [9] is one of popular low-level features of images used for pattern retrieval. The paper presents a proposal that uses the Gabor wavelet features for texture analysis and provides a comprehensive experimental evaluation. Manjunath and Ma [13] indicate the analysis using Gabor wavelet features are more accurate in pattern retrieval than analyses using three other multiscale texture features: pyramid-structured wavelet transform (PWT), tree-structured wavelet transform (TWT), and multi-resolution simultaneous autoregressive model (MR-SAR) features by comparing them.

Another popular feature used to retrieve images from digital image libraries or multimedia databases is color histograms. It is the efficient and insensitive method, but course characteristics as well. So images, which have totally different appearances can have similar histograms. Pass and Zabih [14] propose a *Histogram refinement* technique for comparing images using additional constraints. The technique includes of splitting the pixels in a given bucket into several classes, based upon local property. The pixels in the same class can be compared with others in the same bucket.

3 The Proposed Two-Level Multi-module Image Classification System

In general, the accuracy of the sports image classification depends on the number of sports classes when a single neural network classifier (NNC) is used. As the number of classes increases, there is high possible that the accuracy of the classification decreases. Especially, it happens to have seriously bad performance for a few classes due to he black-box style learning of the neural network. To improve the accuracy of those classes, we propose the two-level multi-module image classification system as shown in Fig. 1. In the figure, the neural network classifier at level 1 identifies one of groups of sports image classes using the input features extracted from the MPEG-7

Color Layout descriptor. The input features are parsed to numerical values, which are suitable for neural network implementation and normalized to the 0-1 range. By normalizing the input features, it can avoid that input features with big number scale dominant the output of the NNC over input features with small number scale.

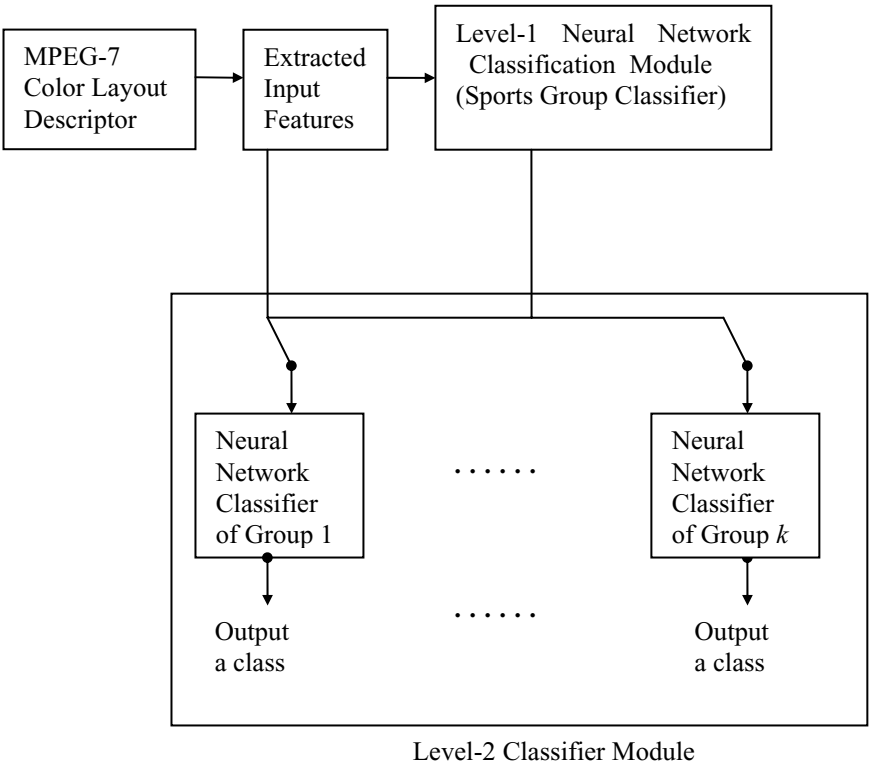

Fig. 1. Two-level multi-module sports image classification system

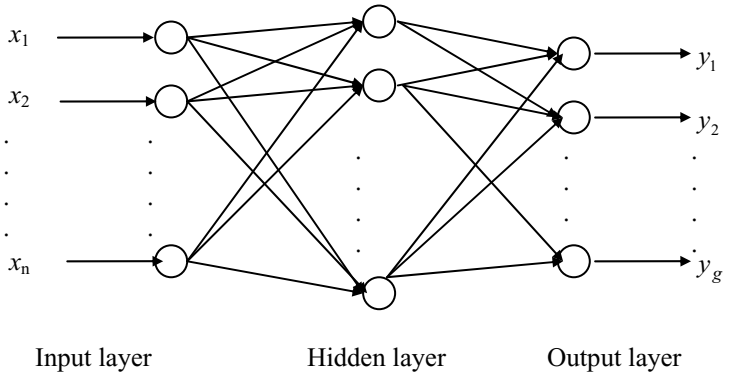

Fig. 2. An example of three layered NNSGC

Using the normalized input features and the groups of sports classes, we can model a neural network sports group classifier (NNSGC). Fig. 2 shows an example of the NNSGC with three layers, i.e., one input layer, one hidden layer, and one output layer.

Let us denote the input feature vector obtained from the MPEG-7 descriptor as $X = (x_1, x_2, \cdots, x_i, \cdots, x_n)$, here x_i is the i^{th} input feature extracted from MPGE-7 Color Layout Descriptor and the subscript n is the dimension of the input features. Also, the output vector for the leve-1 classifier can be expressed as $Y = (y_1, y_2, \cdots, y_i, \cdots, y_g)$, here y_i is the output from the i^{th} output node and the subscript g is the number of sports groups. By utilizing the *hard limit* function in the output layer, we can have binary value, 0 or 1, for each output node y_i as Equation (1).

$$y_i = f_o(netinput_o) = \begin{pmatrix} 1, & netinput_o \geq 0 \\ 0, & \text{otherwise} \end{pmatrix} \tag{1}$$

where f_o is the hard limit function at the output node and $netinput_o$ is the net input of f_o. As shown in Equation (2), the net input is can be expressed as the product of the output vector in the hidden layer, denoted as Y_h, and the weight vector W_o at the output layer.

$$netinput_o = W_o{}^T Y_h \tag{2}$$

With the same way, the hidden layer output vector, Y_h, can also be computed by functioning the product of the input weight vector and the input vector. Thus, the accuracy of the NNSGC depends on the values of whole weight vectors. To obtain the optimal weight vectors, the NNSGC is trained using the back-propagation algorithm which is commonly utilized for training neural networks. The training is done after coding each group of sports classes into g dimension orthogonal vector. For example, if we have two groups of sports classes then each group is coded to (1, 0) or (0, 1).

If the output of NNSGC is group j then the same input features as used in the NNSGC are fed into the corresponding neural network sports class classifier (NNSCS) j at level-2 module to classify the sports classes included in the sports group j. The structure of each NNSCS is almost the same as that of the NNSGC, except the number of layers and nodes. The training for the NNSCS is done with coding each class of sports in the identified group into c dimension orthogonal vector, here c is the number of sports classes in the sports group j.

By using the two-level multi-module sports image classification system, we can reduce the number of sports classes at the classification stage (at level-2). For example, if we have 16 classes with 4 groups then the single classifier needs 16-dimensional output vector. If our system is used, the dimension of output vector reduces to 4. However, the overall accuracy of our system depends on the accuracy of the NNSGC, as seen in Fig. 1.

4 Experiment

4.1 Experimental Environment

We implemented the two-level multi-module sports image classification system using 8 sports image data such as Taekwondo, Field & Track, Ice Hocky, Horse Riding, Skiing, Swimming, Golf, and Tennis. As explained in the previous section, we extracted 12 input features from query images using four MPEG-7 descriptors of Color Layout. The images were classified into two groups according to the characteristics of input features; one group includes Taekwondo, Field & Track, Ice Hocky, Horse Riding and the other group is the rest of classes. Also, 800 samples of were collected, 100 samples per class. For training the NNSGC at level-1 and the NNSCS at level-2, 80 samples per class are used, while 20 samples per class for test. The training and testing images are exclusive. We structured the three-layered NNC for each module. The hyperbolic tangent sigmoid function and hard limit function was used in the hidden layer and in the output layer, respectively. For training the NNC, we chose the back-propagation algorithm because of its training ability. In order to optimal weight vectors, large number of iterations (500,000 in this experiment) is selected.

4.2 Experimental Results

In this section, we compared the performances of the two-level multi-module sports image classification system and the single classification system for test samples. The tables from 1 to 3 show the performance of each module in our system.

Table 1 shows the performance of the level-1 module for identifying sports group. From the table, the accuracy of identifying each group is not perfect but acceptable, 90.79 % for the group 1 and 89.33 % for the group 2.

Table 2 and 3 show the accuracies of the first and the second module at level 2, respectively. From Table 2 and 3, the thing we should not overlook is that the values in the tables are the accuracies with assumption of the perfect identification at level 1. The overall accuracies can be obtained by the product of the accuracy of the identification of each group in Table 1 and the accuracies of the classes included in the group as seen in Table 2 and 3.

Table 1. The accuracies of the level-1 module for identifying sports group (%)

	Sports Group1(Taekwondo, Field & Track, Ice Hockey, Horse Riding)	Sports Group2 (Skiing, Swimming, Golf, Tennis)
Sports Group1 (Taekwondo, Field & Track, Ice Hockey, Horse Riding)	**90.79**	9.21
Sports Group2 (Skiing, Swimming, Golf, Tennis)	10.67	**89.33**

Table 2. The accuracies of the first level-2 module for classifying the classes in group 1 (%)

	Taekwondo	Field & Track	Ice Hockey	Horse Riding
Taekwondo	**88.89**	0.00	5.56	5.56
Field & Track	0.00	**72.22**	11.11	16.67
Ice Hockey	5.56	5.56	**77.78**	11.11
Horse Riding	0.00	0.00	0.00	**100.00**

Table 3. The accuracies of the second level-2 module for classifying the classes in group 2 (%)

	Skiing	Swimming	Golf	Tennis
Skiing	**90.91**	0.00	0.00	9.09
Swimming	4.55	**81.82**	4.55	9.09
Golf	4.55	4.55	**86.36**	4.55
Tennis	9.52	4.76	14.29	**71.43**

Table 4 shows the comparison between the overall accuracies of our system and those of the single classification system. From the table, it is shown that our system is effective for improving the performance of a couple of classes with bad performance when the single classification system is used. For example, the accuracies of Swimming and Tennis are 66.67% and 44.44 %, respectively when the single classification system is used. By utilizing our system, their accuracies are improved to 73.09% and

Table 4. The comparison of overall accuracies of our system and single classification system (%)

		Taek-wondo	Field & Track	Ice Hockey	Horse Riding	Skiing	Swim-ming	Golf	Tennis
Taek-wondo	S	**77.78**	11.11	5.56	0.00	0.00	0.00	0.00	5.56
	M	**80.69**	0.00	5.05	5.05	9.21			
Field & Track	S	0.00	**66.67**	5.56	16.67	0.00	0.00	0.00	11.11
	M	0.00	**65.57**	10.09	15.13	9.21			
Ice Hockey	S	0.00	11.11	**72.22**	0.00	5.56	5.56	0.00	5.56
	M	5.05	5.05	**70.61**	10.08	9.21			
Horse Riding	S	0.00	0.00	5.56	**83.33**	0.00	11.11	0.00	0.00
	M	0.00	0.00	0.00	**90.79**	9.21			
Skiing	S	0.00	0.00	5.56	5.56	**83.33**	5.56	0.00	5.56
	M	10.67				**81.21**	0.00	0.00	8.12
Swim-ming	S	5.56	5.56	5.56	0.00	11.11	**66.67**	0.00	0.00
	M	10.67				4.06	**73.09**	4.06	8.12
Golf	S	16.67	5.56	0.00	5.56	0.00	0.00	**72.22**	0.00
	M	10.67				4.06	4.06	**77.15**	4.06
Tennis	S	11.11	11.11	0.00	5.56	16.67	11.11	0.00	**44.44**
	M	10.67				8.50	4.25	12.77	**63.81**

63.8%, respectively. However, it is shown that there is no big difference on the performances for the classes with acceptable performance for the single classification system. From the result, it can be said that our system can be effective for improving the classification performance of the sports classes with bad performance in the single classification system.

5 Conclusion

This paper proposed the two-level multi-module image classification system for classifying sports images using the input features extracted from the MPEG-7 Color Layout descriptor. As seen in the experimental section, our system can outperform the single classification system for the sports classes with bad performance in the single classification system. In addition, our system does not degrade the performance of the other classes with the acceptable performance in the single classification system. In the future work, different MPEG-7 descriptors will be applied for the classification task. Also, the different combinations of modules and levels are simulated for performance improvement.

References

1. R.A. Jacobs, M.I. Jordan, and A.G. Barto: Task decomposition through competition in a modular connectionist architecture: The what and where vision task, Cognitive Science, 15:219--250, 1991.
2. A. Waibel, H. Sawai, and K. Shikano: Modulraity and scaling in large phonemic neural networks. IEEE Transaction on Acoustics, Speech and Signal Processing, 37:1888-1897, 1989
3. D.L. Reilly, C. Scofield, C. Elbaum, and L.N. Copper: Learning system architectures composed of multiple learning modules. In IEEE Int'l Conf. Neural Networks, p.p. 495-503, 1989
4. R.A. Jacobs and M.I. Jordan: Hierarchical mixtures of experts and the EM algorithm. Neural Computation, 6:181-214, 1994
5. F.J. Smieja: Multiple network systems Minos modules: Task division and module discrimination, Proc. Of the 8th AISB conference on Artificial Intelligence, Leeds, April 1991
6. L.Y. Pratt and C. A. Kamm: Improving phoneme classification neural network through problem decomposition. In IEEE Int'l Joint Conf. Neural Networks, p.p. 821-826, 1991
7. D.N. Osherson, S. Weinstein, and M. Stoli: Modular learning, Computational Neuroscience, p.p. 369-377, 1990
8. S. Haykins, Neural Networks: A comprehensive Foundation. IEEE Press, Macmillan, NY, 1994
9. M. Gorkani and R. W. Picard, "Textual orientation for sorting photos at a glance," In Proceedings of International Conference on Pattern recognition, pp. 459-464, Jerusalem, Israel, October 1994
10. A. Vailaya, A. Jain, and H. J. Zhang, "On image classification: city images vs. landscapes," In Proceedings of IEEE Workshop on Content – Based Access of Image and Video Libraries, December 1998

11. M. Szummer and R. W. Picard, “Indoor-outdoor image classification,” In Proceedings of IEEE International Workshop on Content – Based Access of Image and Video Libraries, Bombay India, December 1998
12. Y-I Ohta, T. Kanade, and T. Sakai, “Color information for region segmentation,” Computer Graphics and Image Processing, vol. 13, No. 3, pp. 222-241, July 1980.
13. B. S. Manjunath and W. Y. Ma, “Texture features for browsing and retrieval of image data,” IEEE Transactions on Pattern Analysis and Machine Intelligence,” vol. 18, no. 8, August 1996.
14. G. pass and R. Zabih, “Histogram refinement for content-based image retrieval,” In Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision, Sarasota, Florida, USA, December 1996.