Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Efficient Entropy-based Features Selection for Image Retrieval

Tsun-Wei Chang
Department of Computer Science and Information Engineering
De Lin Institute of Technology
Tucheng, Taipei County, 236 Taiwan
e-mail: alan1107@dlit.edu.tw

Yo-Ping Huang
Department of Electrical Engineering
National Taipei University of Technology
Taipei, 10608 Taiwan
e-mail: yphuang@ntut.edu.tw

Frode Eika Sandnes
Faculty of Engineering
Oslo University College
Oslo, Norway
e-mail: frodes@hio.no

***Abstract*—Information retrieval systems should provide users quick access to desired information. There are no established ways for inexperienced users to explicitly express queries for retrieving images from ecological databases. This study proposes an entropy-based feature selection strategy for finding images of interest from databases. Six visual features are used to represent birds, and hence used to formulate search queries. The proposed method is tested on a real world bird database and the experimental results demonstrate the effectiveness of the presented work.**

***Keywords*—entropy, content-based image retrieval, feature selection**

I. INTRODUCTION

Retrieving information of interest from an image database is an important issue in many applications and an open research problem. Traditional keyword-based queries can assist users in finding information in which they are interested. This assumes that users know exactly what they want. Intuitiveness and straightforward expressiveness are the two main characteristics of the text-based image retrieval approach. However, text-based queries cannot be used for more elaborate queries. For many data types, it is difficult to conduct a query with a complete set of keywords. Moreover, even if images are characterized and annotated, difficulties still occur because the expression of a perception is individualistic and not necessarily uniform. Consequently, the retrieval results are often unsatisfactory. For example, a non-expert bird watcher may not confidently describe the features of a bird that they have seen. Moreover, expert bird-watchers may possibly spot different details because they may have different focus and emphasis. Due to the inexact and uncertain nature of textual descriptions, researchers have attempted to find better and more sophisticated image retrieval strategies. Content-based image retrieval (CBIR) technology is one such an approach.

CBIR has been an active research area during the last decade. In CBIR systems, an image is represented by a set of low level visual features [1-4]. The image retrieval can then be achieved by comparing these low level features by computing the similarity of the features. One problem is that there is a gap between high-level concepts and low-level features. This gap hinders the further development of CBIR systems. To bridge this gap, region-based image retrieval (RBIR) technology has been proposed. The RBIR approaches segment images into several regions and the low level feature are extracted from these regions [5-7]. Images can be presented on the object level based on region features. Consequently, human perception can be better fitted compared to global features that are extracted from entire images.

Features are the foundation for retrieving images with both the CBIR and RBIR techniques. Therefore, the focus is on finding suitable features to represent an image. An increasing number of features are used to build image retrieval models for efficiently retrieving images. Consequently, the high dimension feature vectors lead to complex and time-consuming computations. As CBIR and RBIR technologies mature, it is appropriate to address the problem of feature selection. A feature selection criterion concerns the process of selecting a subset from a set of original features. Different image retrieval objectives will lead to different feature subset needs [8-11]. The choice of a suitable volume of features needs to be cautiously considered. Moreover, the design of an intuitive user interface that captures these features is also a key challenge. To solve these problems, an efficient entropy-based feature selection strategy is employed in the image retrieval system. A bird image dataset is used to demonstrate the scheme.

Our approach involves visual features characterization and knowledge representation. A major component of our approach is the bird information features consisting of several types of concepts. Visual characterization concepts addressing a database entity are viewed as the primary features for retrieving bird information. In addition, keyword-based search is also provided for users accustomed to this mode of access. Three frameworks are combined in the proposed system including: (1) a knowledge acquisition phase, (2) a primary search phase and (3) a key-word search phase. A major issue is the precise query when retrieving bird information from the database based on the visual features. Users may not know the scientific name of a bird. Consequently, the query may not match the expected search result. Impatient users also demand short response times. We utilize six visual features to characterize a bird image in the database and exploit entropy-based feature selection. In addition, the system finds the most relevant candidates through the information ontology. In addition to identifying a perfect match, the most relevant matches are retrieved according to the bird information ontology. This is especially helpful for vague queries where the users do not know how to specify the queries.

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

The remainder of this paper is organized as follows. Section II reviews the entropy measurements from a technical perspective. Section III illustrates the proposed system architecture. Section IV presents the experimental results. The conclusions and future research issues are presented in section V.

II. Entroy Measurement

Entropy is a measurement of the degree of uncertainty that exists in a system. While observing the outcome of a random experiment, a measurement of the information can be obtained. This concept has been used in many fields including information theory, mathematics, statistics, and economics [12-14]. Of these applications, Shannon contributed the broadest and the most fundamental definition of the entropy measurement [15]. Shannon's entropy is an important measure for evaluating structures and patterns in the data. The lower the entropy (uncertainty) the more structure is given in the relation. The entropy is first introduced. Then, feature selection based on the entropy is proposed.

A. Entropy Computation

Shannon's Entropy is used to measure the uncertainty of a random variable X which takes different probabilities among a set of finite values into account. Let X be a random variable with a finite set of values containing n symbols given by $\{x_1, x_2, ..., x_n\}$ and P the space of all possible probability distributions. If a specific value x_j occurs with probability distribution $p(x_j)$ such that $p(x_j) \geq 0, j = 1,2,...,n$ and $\sum_{j=1}^{n}(x_j) = 1$, then the information amount associated with the known occurrence of output x_j is defined as follows.

$$I(x_j) = -\log_2 P(x_j)\ . \qquad (1)$$

That means the information generated in selecting symbol x_j is $-\log_2 p(x_j)$ bits for a discrete source. On average, if the symbol x_j is selected $n \times p(x_j)$ times in n selections, the average amount of information obtained from n source outputs is as follows.

$$\begin{aligned} &-n\times p(x_1)\log_2(x_1) - n\times p(x_2)\log_2(x_2) - \cdots \\ &- n\times p(x_n)\log_2(x_n). \end{aligned} \qquad (2)$$

The entropy is such formulated as a function of the distribution of random variable X which will rely on the probabilities. Hence, entropy $E(X)$ is the average information, and is defined as follows.

$$E(X) = -\sum_{j=1}^{n} p(x_j)\log_2 p(x_j)\ . \qquad (3)$$

B. Entropy-based Feature Selection

When retrieving bird images from a photographic database, a high dimensional feature space will affect the computation and storage complexity. Much research has addressed the problems of high dimensional feature spaces and feature selection. Some methods compute a score for each feature and then select the feature which gets the highest scores when retrieving images. We utilized the entropy derived from its probability to compute the feature score. Assuming that the values are described with a feature vector $\{x_1, x_2, ... x_n\}$, where x_i is the value of the i-th feature f_i. In our case, the feature type is symbolic and the probability of feature f_i is defined as follows.

$$p(x_i) = \frac{\#\text{ of queries with } f_i}{\text{total number of queries}}\ . \qquad (4)$$

The retrieval process is logged in the database such that a sample of queries can be used to derive the probability.

When a query q with feature f_i is issued in a retrieval round, the entropy $H^i_{R_q}$ of feature f_i can then be computed on this query result set R_q as follows.

$$H^i_{R_q} = -\sum_{x_i \in X} p^i_{x_i} \log(p^i_{x_i})\ . \qquad (5)$$

Where x_i is the domain of f_i, and $p^i_{x_i}$ is the probability of observing the value x_i in R_q. As we know that a feature with larger entropy is more likely to reduce the result set if it is constrained.

The entropy-based features elimination starts with the full feature set. The features with higher entropy are less relevant to the retrieval target, and they will be removed from the candidate features. As a consequence, removing most irrelevant features can reduce the computation time in retrieving images.

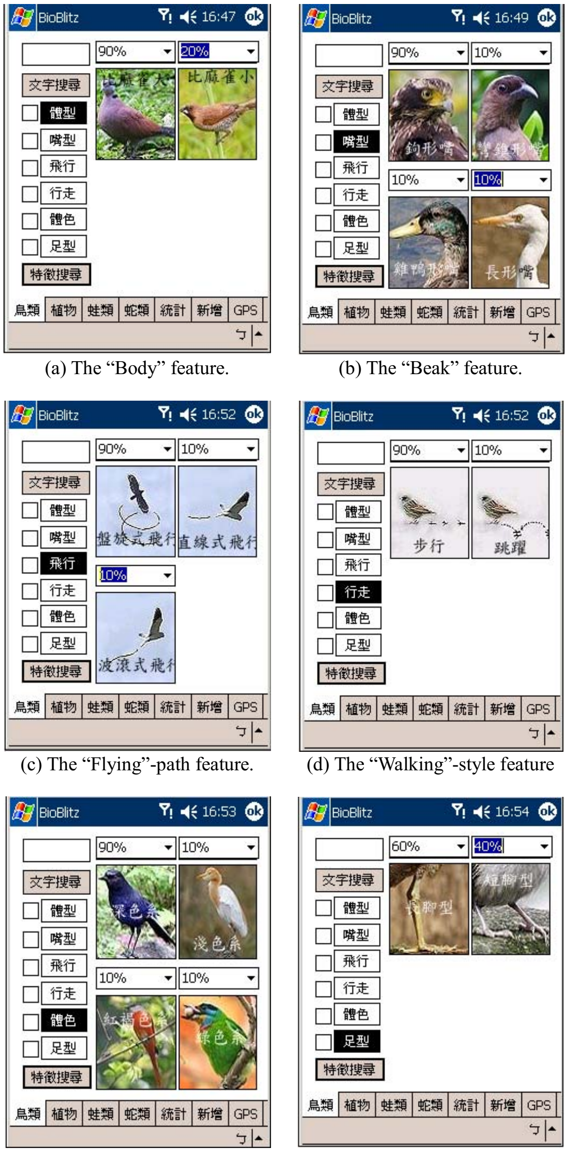
III. System Implementation

We proposed an enhanced multimedia information retrieval method. The approach enables users to express their queries in terms of both simple visual features and keywords. Moreover, the retrieval efficiency is further improved via the entropy-based feature selection that is constructed from the diversity categorization. The proposed searching methodology is illustrated next.

A. Data Representations

Six commonly observed bird features $(f_i, i = 1,2,...,6)$ and their corresponding optional sub-features, listed in Table 1, are used to represent a database entity or the user query. The users specify their queries by clicking on the respective icons in the data entry form as depicted in Fig. 1.

For any specific feature, users can give a confidence degree based on their impression on that query entity. Then the system transforms the specifications into a value vector $I = \{I_1, I_2, I_3, I_4, I_5, I_6\}$. And the j-th value I_j for its corresponding visual feature f_j could have optional k terms, i.e., options from sub-feature, $2 \leq k \leq 4$. That means the value I_j for some particular visual feature contains its corresponding optional terms in form of $I_j = \{I_{j1}, ..., I_{jk}\}$. A term I_{jp} (p=1,…, *k*) indicates the confidence to which users prefers an optional sub-feature.



(a) The “Body” feature.

(b) The “Beak” feature.

(c) The “Flying”-path feature.

(d) The “Walking”-style feature

(e) The “Color” feature

(g) The “Foot” feature

Figure 1. Six major features and their corresponding optional sub-features.

For example, a user watches a large flying bird and sets the values of the 1st visual feature to $I = \{0.7, 0.3\}$. This reveals that the user has a 70% confidence of seeing a bird with big body and 30% confidence of a small one shown in Fig. 2. Similarly, a user can give different values for all the observed features. Thus, the value vector I can be formulated to represent the query object according to the user’s perception.

TABLE I. BIRD FEATURES USED IN THE SYSTEM.

Visual Feature	Sub-feature values (Options for each feature)
Body size (f_1)	Similar to or smaller than a sparrow (I_{11}), Bigger than a sparrow (I_{12}).
Beak shape (f_2)	Hooked type (I_{21}), Bradawl-like type (I_{22}), Duck-like type (I_{23}), Long type (I_{24}).
Flying style (f_3)	Spiral (I_{31}), Straight (I_{32}), Wave and roll. (I_{33})
Walking style (f_4)	Walk (I_{41}), Jump (I_{42}).
Feather color (f_5)	Dark (I_{51}), Light (I_{52}), Red-brown (I_{53}), Green (I_{54}).
Foot length (f_6)	Long (I_{61}), Short (I_{62}).

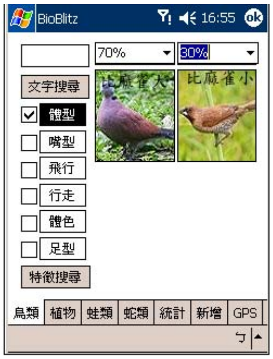


Figure 2. An example in specifing the confidence to the corresponding sub-features.

B. Information Retrieval Strategy

The database entities are predefined for the corresponding optional sub-features of each visual feature. A “1” is given if the optional term is true; otherwise a “0” is assigned for that term. Each entity in the database is transformed into a vector I' containing the confidence degree for the corresponding optional term.

When a user issues a query with different level of degrees for the optional terms, the query vector Q is formed as $Q(I)$, where I is a value vector for the query object. The target object in the database is represented as $T(I')$, where the I' is also a value vector for a database entity. Each term in the vector equals the probability of a correct guess if a user is unfamiliar with the associated feature. The dissimilarity measure (distance D) is computed to capture what the user observed:

$$D(Q,T) = d(I, I'), \tag{6}$$

$$d(I, I') = \sqrt{\sum_{j=1}^{6} \frac{1}{H^j} \times \sum_{k=1}^{p} (I_j[k] - I'_j[k])^2}, \tag{7}$$

where p is the number of optional sub-features for the j-th visual feature and H^j is the entropy of the corresponding visual feature. A visual feature with larger entropy indicates a higher importance of this feature. The Euclidean distance is used to quantify the dissimilarity $d(I, I')$ between a query and instances in the database.

C. Keyword Queries

Since keyword-based queries are still the most widely used search paradigm, the proposed system also allows users to express their queries by a set of keywords. In the last decade, ontologies have played an important role in knowledge-based system. Depending on the level of generality, any information structure can be called an ontology, including a table of contents, an introduction to an article, metadata, category tree structure or database. There is not a widely accepted classification of ontology types, but certain ontology types are more common than others. Here, the domain-task ontology is exploited in the information retrieval system. The domain-task ontology specifies the required vocabulary and knowledge in a given domain that are used to solve problems associated with a specific task.

The ontology provides a vocabulary, or textual terms, and their relations to model the domain knowledge. The proposed system supports multi-concept searches for target information. If the primary match (visual feature query) fails to yield the desired results, the proposed system performs another advanced search based on the text query through the ontology to widen the search. Unlike traditional information retrieval systems where queries may return unsatisfactory answers or even no answers, our system finds the most probable matches according to the user's specification.

Both the scientific name and popular name of a bird can be searched using text. The system guesses the bird name even if the name is only partially provided. An example is that a user might not know the exact name of a "Green-winged Teal" and it will be very difficult for one to retrieve its information in terms of a query of indirect keywords under traditional keyword matching manner. In our system, a user can input only a word "鴨" in Chinese or the exact name "小水鴨 (Green-winged Teal in English)" to retrieve the related information of that bird. Of course, a user can also input the formal scientific name "Anas crecca" as the query for "Green-winged Teal".

The visual features allow user to make more intuitive queries. Inexperienced users may specify a query based on what they have seen or their impression of a bird. A query is formed through simple GUI.

The bird habitat place domain knowledge is useful for bird watchers. In most cases, when the bird watchers observe birds around some wetland, they may want to investigate or count the bird species in some particular outdoor observational areas. For example, bird watchers that visit the Guandu Nature Park or the Hua-Jian Wild Duck Natural Park in northern Taiwan may be concerned about the number of species dwelling in that area. The user can input the keyword "Guandu" or "Hua-Jian" to retrieve information about the bird habitat in this area. The current database comprises 183 Taiwanese species from 43 families. The grouping of birds into different categories help users retrieve advanced information and it also has educational benefits.

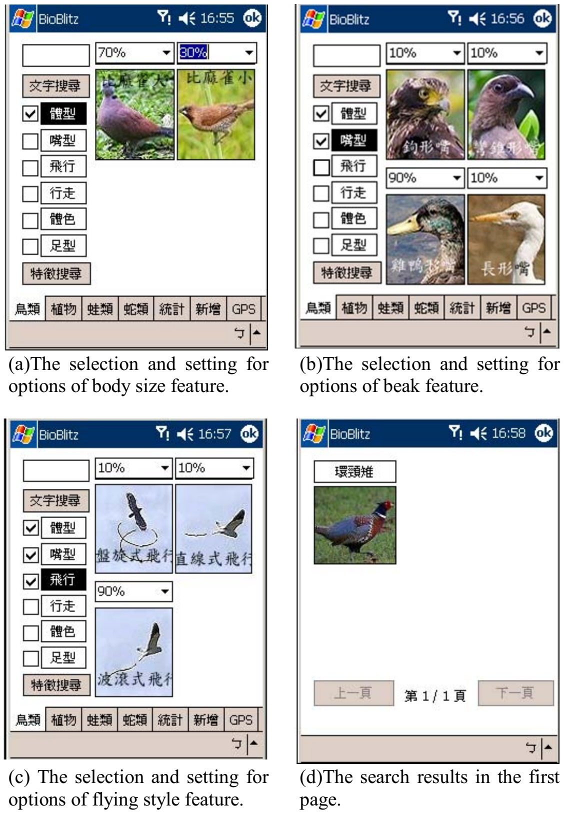
Users can add new vocabulary into the ontology using their individual knowledge about the birds. The annotations are stored in the repository. Whenever users see a bird, they can easily note down the information about the observed bird by using the life note function. The user may have valuable knowledge about the bird that can be shared with others through the sharable bird ontology.

IV. Experimental Results

The proposed bird information retrieval system is developed using eMbedded Visual C++ 4.0 and runs in a Windows environment on a Pocket PC. In this study, we built an ecological information data model to deal with the bird information retrieval task. The implementation focuses on data information representation, entropy-based feature selection, information-based ontology construction and similarity measurements. Currently, a collection of 183 bird species in Taiwan and their related information are stored in the database. The database entities belong to 42 families (categories) among 217 families in Taiwan. There are currently 481 bird species belonging to 217 families in Taiwan. Our collection includes 38% of Taiwan's bird species. The collection is continuously increasing. In addition, the database contains bird sounds, images, text descriptions and life notes. Six visual features are used to capture the characteristics of an observed bird that it is difficult to describe using a set of keywords. Through the proposed system, the users can easily specify what they have seen and issue a query. The intuitive visual feature entry screens provide users with a flexible user-friendly interface. The preliminary experimental results show that the proposed system is a feasible framework for dealing with such a semantic retrieval task for an ecological database.

An example is demonstrated in Fig. 3. A vector $I_1 = \{0.7, 0.3\}$ denotes the user has 70% of confidence that the observed bird is bigger than a sparrow shown in Fig. 3(a). Similarly, a vector $I_2 = \{0.1, 0.9, 0.1, 0.1\}$ means the user has 90% confidence in the Duck-like type beak and 10% confidence in the hooked beak for a given instance shown in Fig. 3(b). Meanwhile, a vector $I_3 = \{0.1, 0.1, 0.9\}$ means the user has 90% confidence in the wave and roll flying style, 10% confidence in the spiral flying style and 10% confidence in the straight flying style for a given instance shown in Fig. 3(c).

Together with the value of the term assigned by the users and predefined by the system, the dissimilarity measure is computed from the value of an entity in the database. After the selection of options for each visual feature, the query is then represented by a vector. The primary search results will then be displayed on the screen shown in Fig. 3(d).



(a)The selection and setting for options of body size feature.

(b)The selection and setting for options of beak feature.

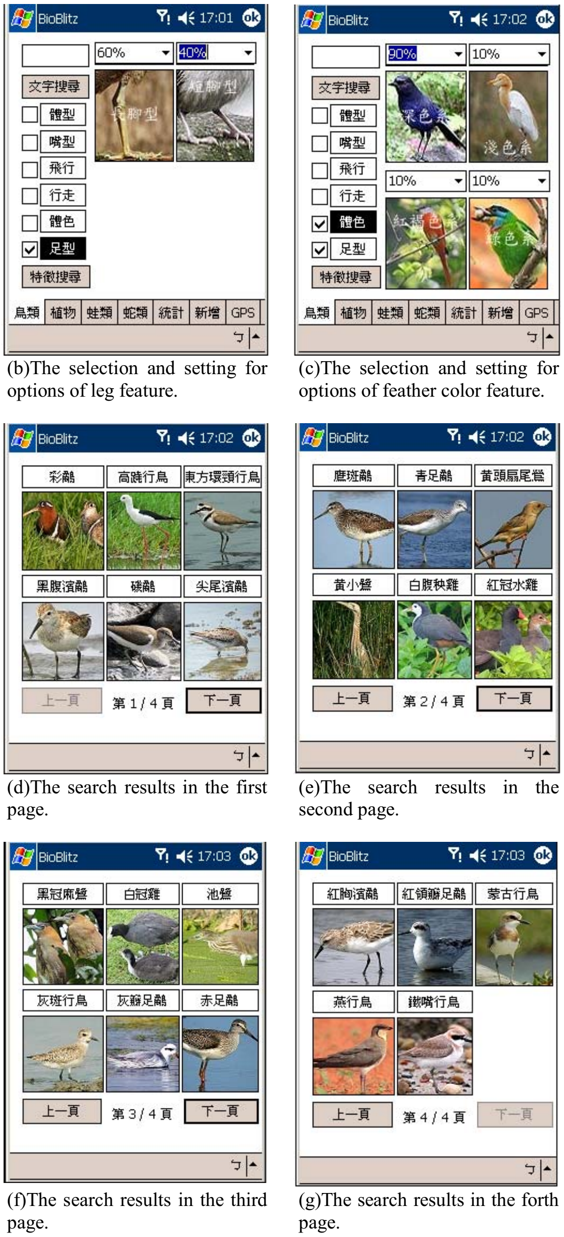
(c) The selection and setting for options of flying style feature.

(d)The search results in the first page.

Figure 3. The primary search results based on selection of three main visual features.

Another example is demonstrated in Fig. 4. The user has watched the bird shown in Fig. 4(a) and he/she set a vector I_6 = {0.6, 0.4} which denotes the user only has 60% confidence in the observed bird with long legs shown in Fig. 4(b). Nevertheless, he/she is 90% confident that the feather color is dark for the given instance such that the vector I_5 will be {0.1, 0.9, 0.1, 0.1, 0.1} shown in Fig. 4(c). The retrieval results are shown in Fig. 4(d) to Fig. 4(g). The target bird is located at rank 3.

(a) An example bird observed by a user.



(b)The selection and setting for options of leg feature.

(c)The selection and setting for options of feather color feature.

(d)The search results in the first page.

(e)The search results in the second page.

(f)The search results in the third page.

(g)The search results in the forth page.

Figure 4. Additional search results when using (a) as the query.

Furthermore, if the current match is unsatisfactory, the proposed system performs another search based on the query using the bird information ontology to get additional related information. Fig. 5 shows the top 6 search results when using the keywords "Family Anatidae". The search is performed through the proposed information-based ontology and the results demonstrate the feasibility of the proposed system.

Unlike traditional information retrieval systems in which implicit queries may return unsatisfactory answers or even no answer, the proposed system always retrieves related matches according to what the users specify.

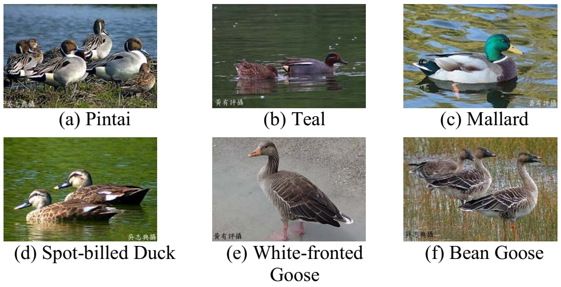

(a) Pintai
(b) Teal
(c) Mallard
(d) Spot-billed Duck
(e) White-fronted Goose
(f) Bean Goose

Figure 5. The top 6 search results while using the “Family Anatidae” as query term.

The pocket PC implementation has been demonstrated at the first BioBlitz Biodiversity Festival held at Hua-Jian Wild Duck Natural Park for bird watchers.

V. Conclusions and Future Work

In this paper, a bird retrieval strategy is proposed where a visual feature query is used as the fundamental search mechanism. We did not intend to build a complex data model. Instead, a handy bird information retrieval system with simple interface is proposed. The integration of entropy-based visual feature selection query and bird information ontology contributes the searching effectiveness.

Since several descriptors are used simultaneously, it is necessary to integrate similarity scores resulting from the matching processes in different feature spaces. The entropy is used to adjust the uncertainties of users. It is potentially useful when the user cannot be reasonably sure about this feature. With the assistance of the proposed user interface, he/she can easily describe the query formulation.

This experimental system has proved to be generic and flexible, although there is room for improvements. Future work involves investigating more relevant information for the ecological database. Moreover, how should a sharable information ontology be provided and at the same time avoid information overloading. Next, we will derive a more elaborate intensity vector which considers an instance belonging to different classes. The current version does not consider that a bird may belong to multiple categories.

Acknowledgment

This work is supported by National Science Council, Taiwan under Grant 97-2221-E-027-034-MY3.

References

[1] D.C. Young, C.K. Nam and I.H. Jang, “Content-based image retrieval using multiresolution color and texture features,” *IEEE Trans. on Multimedia*, Vol. 10, Ixxue 6, pp.1073-1084, Oct. 2008.

[2] A. Grigorova, F.G.B.De Natale, C. Dagli and T.S. Huang, “Content-based image retrieval by feature adaptation and relevance feedback,” *IEEE Trans. on Multimedia*, Vol.9, Issue 6, pp.1183-1192, Oct. 2007.

[3] J. Yu and Q. Tian, “Semantic subspace projection and its applications in image retrieval,” *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 18, Issue 4, pp.544-548, Apr. 2008.

[4] R. Rahmani, S.A. Goldman, H. Zhang, S.R. Cholleti and J.E. Fritts, “Localized content-based image retrieval,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 30, Issue 11, pp.1902-1912, Nov. 2008.

[5] P. Mylonas, E. Spyrou, Y. Avrithis and S. Kollias, “Using visual context and region semantics for high-level concept detection,” *IEEE Trans. on Multimedia*, Vol. 11, pp.229-243, Issue 2, Feb. 2009.

[6] W. Jiang, G. Er, Q. Dai and J. Gu, “Similarity-based online feature selection in content-based image retrieval,” *IEEE Trans. on Image Processing*, Vol. 15, Issue 3, pp.702-712, March 2006.

[7] F. Li, Q. Dai, W. Xu and G. Er, “Multilabel neighborhood propagation for region-based image retrieval,” IEEE Trans. on Multimedia, Vol. 10, Issue 8, pp.1592-1604, Dec. 2008.

[8] M. Vasconcelos and N. Vasconcelos, “Natural image statistics and low-complexity feature selection,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 31, Issue 2, pp.228-244, Feb. 2009.

[9] K. Huang and S. Aviyente, ”Wavelet feature selection for image classification,” *IEEE Trans. on Image Processing*, Vol. 17, Issue 9, pp.1709-1720, Sep. 2008.

[10] P.E. Meyer, C. Schretter and G. Bontempi, “Information-theoretic feature selection in microarray data using variable complementarity,” *IEEE Journal of Signal Processing*, Vol. 2, Issue 3, pp.261-274, June 2008.

[11] S.-Y. Kung and M.-W. Mak, “Feature selection for self-supervised classification with applications to microarray and sequence data,” *IEEE Journal of Signal Processing*, Vol. 2, Issue 3, pp.297-309, June 2008.

[12] D. Farina, E.N. Kamavuako, J. Wu nad F. Naddeo, “Entropy-based optimization of wavelet spatial filters,” *IEEE Trans. on Biomedical*, Vol. 55, Issue 3, pp.914-922, March 2008.

[13] M.S. Manikandan and S. Dandapat, “Multiscale entropy-based weighted distortion measure for ECG coding,” *IEEE Signal Processing Letters*, Vol. 15, pp.829-832, 2008.

[14] D. Sen and S.K. Pal, ”Generalized rough sets, entropy, and image ambiguity measures,” *IEEE Trans. on Systems, Man, and Cybernetics*, Part B, Vol. 39, Issue 1, pp.117-128, Feb. 2009.

[15] C.E. Shannon, “A mathematical theory of communiction,” *The Bell Systems Technical Journal*, Vol. 27, pp.379-423, 1948.