Image Retrieval and Feature Vector Size Reduction

Hossein SadeghianNejad, Jamshid Shanbehzadeh, Abdolhossein Sarafzadeh.

***Abstract*—This paper focuses on the effectiveness of image features on retrieval performance when searching for similar photos. The nature of image features with respect to their relevancy to the similarity measure can heavily affect retrieval rate. Retrieval systems can exploit these characteristics to reduce the number of features considered and select the most effective features for efficient content-based image retrieval. This phenomenon results in the improvement of retrieval performance in terms of speed, recall rate, precision rate and F-Measure. This paper employs the F-norm of wavelet coefficients for the feature vector. This algorithm consists of extracting wavelet coefficients of images and finding the F-norm for the image features, and then comparing the F-norm of the query image with the F-norm of the images in the database. The ReliefF algorithm selects the most relevant and non-redundant image features for the retrieval process. This algorithm reduces the dimension of the feature vector by half without side effects on performance. This paper uses the Corel [1, 2] dataset image in simulations and recall rate, precision rate and F-measure as the evaluation criteria.**

Key words: Content Based Image Retrieval (CBIR), Feature Extraction and Selection, ReliefF Algorithm, Wavelet Transform

I. Introduction

Content-based image retrieval (CBIR) is currently in high demand because of its wide applications for retrieving images in different areas such as the Internet, medical image archiving, and videos [3, 4, 5]. The development of affordable high quality cameras and the easy distribution of images on shared storage devices have both facilitated the increasing demand for CBIR applications. The basis of CBIR is to express an image in terms of a set of features, and to retrieve similar images by measuring the similarity of these features. There are many types of features that have been employed to construct feature vectors (FV), such as color [6], texture [7, 8], shape [9, 10] and salient points. This paper employs the Frobenius norm [11] of three components of a wavelet transform (WT) for its FV, which is robust against rotation and translation.

The paper focuses on the problem of reducing FV size by selecting the most effective components of the WT coefficients. This is performed using a feature weighting algorithm called ReliefF [12, 13]. This algorithm iteratively estimates feature weights according to their ability to discriminate between neighboring patterns. To do this, we first construct an FV for each image of the Corel dataset using the F-norm of three components of a WT. Then we calculate a weight for each feature using the ReliefF algorithm and select the k top features as the effective subset. We calculate the results of the retrieval and compare them with results of classification and retrieval when we consider all features. The results of the comparisons show that we can reduce the number of features and improve the retrieval performance through selecting prominent features.

The rest of this paper is organized as follows. The next section explains the retrieval algorithm, which consists of a WT, image feature extraction and selection and similarity measures. We then introduce the evaluation criteria and present the experimental results.

Manuscript received on 9th December 2011 and accepted on 23rd December 2011

H. SadeghianNejad is with Tarbiat Moallem University of Tehran, Tehran, IR. Iran (E-mail: h.sadeghiyan@tmu.ac.ir).

J. Shanbehzadeh is an Associate Professor with Tarbiat Moallem University of Tehran, Tehran, IR. Iran (E-mail: jamshid@tmu.ac.ir).

A. Sarafzadeh is an Associate Professor and Head of Department of Computing at Unitec, New Zealand. (E-mail: hsarrafzadeh @unitec.ac.nz).

II. Image Retrieval Algorithm

Figure 1 shows a block diagram of a CBIR system. First, all of the images in the dataset undergo feature extraction: the suitable features in the image are selected and the FV is generated from these selected features. The same procedure is performed for each query image. The FV of each query image is compared with the FVs of the other images in the dataset according to a similarity measure, and the most similar images are selected as the output results.

III. Feature Vector Extraction

The feature extraction consists of producing the wavelet transform coefficients (WTC); the FV components are the norms of the WTC. The most suitable subset of FV components for providing the best retrieval rate is found by the ReliefF algorithm, and is considered to be the final FV. The similarity measure is based on the closeness of the FV components. Next, WT, FV generation and the selection and similarity measures are described.

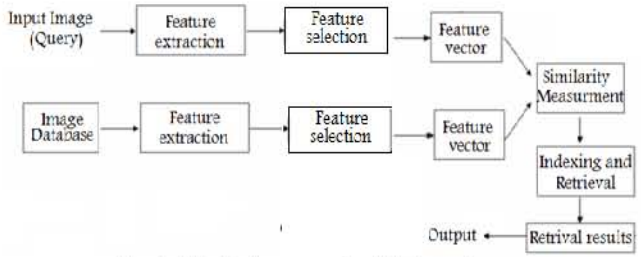

Fig 1. Block diagram of a CBIR system.

The primary FV consists of the norms of three components of a WT. The norms of the coefficients of LL, LH and HL are considered to be the FV components. Equation 1 shows the norm of each wavelet component. We calculate the norm

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

of the rows of the LL and LH versionsof the wavelet coefficients and the columns of the HL version, as shown in Fig 2, Eq. 2 shows the FV obtained by the calculated norms.

$$if\ A \in R^n \Rightarrow \|A\| = (\sum_{i=1}^{n} (a_i)^2)^{\frac{1}{2}} \quad (1)$$

$$V_F = \{ \|LL_1\|, \dots, \|LL_n\|, \|LH_1\|, \dots, \|LH_n\|, \|HL_1\|, \dots, \|HL_n\| \} \quad (2)$$

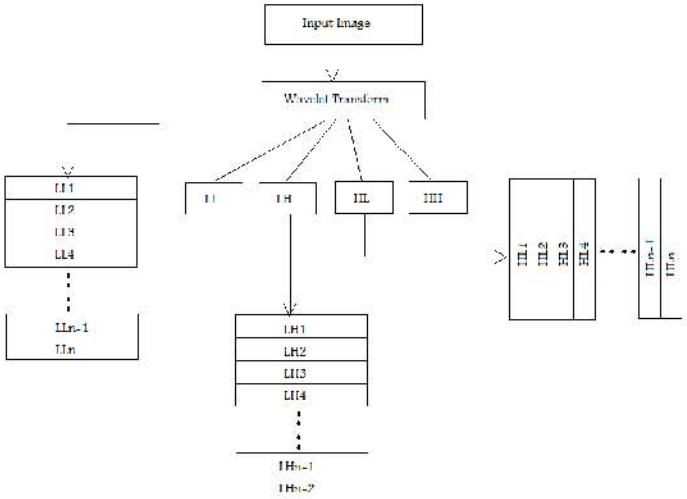

Fig. 2. The process of feature vector generation; three components of wavelet coefficients are used in FV generation.

IV. Feature Vector Selection

Feature selection is the most complicated part of this paper. This refers to the problem of choosing a small subset of features that is suitable to describe the imagefor retrieval applications [14]. Assume that an original feature set has D features. The goal of feature selection is to find the most informative subset of d features, where $d \leq D$. This requires removing irrelevant and redundant features. The motivation behind feature selection in image retrieval is to reduce dimensionality in order to speed up the retrieval algorithm and improve itsoverall performance [15].Another group of feature selection schemes is based on feature weighting rather than feature removal. These algorithms rank features by finding their weight based on a desired cost function. This cost function may be the closeness of images based on a similarity measure.

One well-known feature weighting algorithm is Relief, whose application is in situations with two classes. However, image retrieval often involves hundreds of image classes; therefore an extension of Relief is ReliefF, which is suitable in this condition [16]. This algorithm finds the weight of features by minimizing their inter-class distance and maximizing their intra-class distance. In each step and for each feature, as shown in Figure 3, ReliefF takes the nearest sample in its class and the k nearest samples from the rest of classes; then it updates Equation 3 and repeatsthis process several times until either an unchangeable weight or the end of data is reached.

$$w[i] = w[i] - \frac{diff(i, R, H)}{m} + \sum_{m} [p(c) \times \frac{diff(i, R, M(c))}{m}] \quad (3)$$

The parameters in Equation 3 arethe feature number (i), a randomly selected image (R), the nearest inter-class member (H), the nearest intra-class members (M(C)), the prior probability of each image class (P(C)) and a user defined parameter (m). The detail information on ReliefF can be seen in Figure 4.

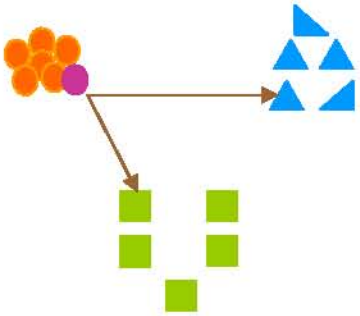

Fig. 4. Multiclass data; the nearest from member of each class is used in ReliefF algorithm.

V. Similarity Measure

Several measures of feature similarity exist, such as Euclidian distance and the angle between FVs [17, 18]. This paper employs a very simple and effective similarity measure based on the average sum of the ratio of the minimum value of an FV over the maximum value. The reason for this is that if the two components are similar, this ratio will be high; otherwise it will be low [19]. Equation 4 shows the similarity measure for one component of a FV. In this equation, v_d^i and v_q^i are respectively the i^{th} component of the FV of the query image and the FV from the image dataset. The average of all S_i is the final similarity factor as shown in Equation 5. The most similar images to the query imageare considered the desired output.

$$S_i = \begin{cases} \frac{\min(v_d^i, v_q^i)}{max(v_d^i, v_q^i)} v_d^i \ or \quad v_q^i \neq 0 \\ 1 \qquad\qquad\qquad\qquad v_d^i, v_q^i = 0 \end{cases} \quad (4)$$

$$S = \frac{\sum_{i=1}^{n} S_i}{n} \quad (5)$$

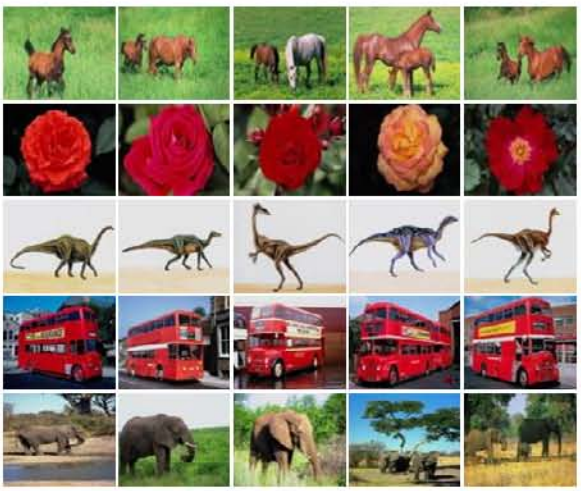

Figure 5: Sample images from a database.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

VI. System Evaluation

Precision rate (PR) and recall rate (RR) are normally employed to evaluate a CBIR system. Equations 8 and 9 show these parameters respectively [20, 21]. In these equations, T and P are the total sets of retrieved images and relevant images based on a desired similarity. Normally for each image all of the members of P are known, and the retrieved subset of P is used in Equations 8 and 9. In the case of the Corel dataset, as all of the images have labels the intersection of T and P can be found without human intervention. The absolute value notation shows the total number of elements in a set.

$$\text{Precision} = \frac{|P \cap T|}{|T|} \tag{6}$$

$$\text{Recall} = \frac{|P \cap T|}{|P|} \tag{7}$$

A suitable retrieval system should have both a high RR and a high PR at the same time. As presented in Equation 8, an F-Measure can be used to show system suitability [22]. This factor has a high value if both RR and PR are high at the same time; otherwise it has a low value.

$$F-\text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

VII. Experimental Results and Future Work

For this research we performed simulations on the Corel database, which contains 1000 labeled images across different categories. This means that the similar images for any given query image are already known. The main focus of this paper is the dimension reduction of FVs without reducing performance. Figures 5 to 7 show three evaluation parameters. These figures show that considering about 50 percent of the featuresin an image gives better performance than considering all of the features. The employment of features in all simulations is based on their rank. The features with higher weights are considered first.

This paper used special features, but the algorithm can be used for any kind of features and this is an open area. As much as we can improve the performance of retrieval via employing sophisticated image features, we can also improve overall performance by using feature selection schemes. A drawback of ReliefF is ignoring features dependency in the removal phase, but there are advanced feature selection methods that consider feature relevancy during the removal phase [23]. Investigating the effects of using these methods in image retrieval would be an interesting topic for future research.

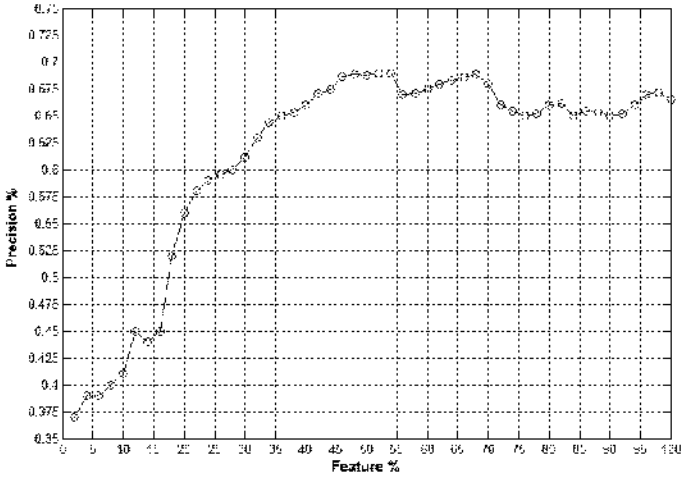

Fig 6. Precision rate based on percentage of features used

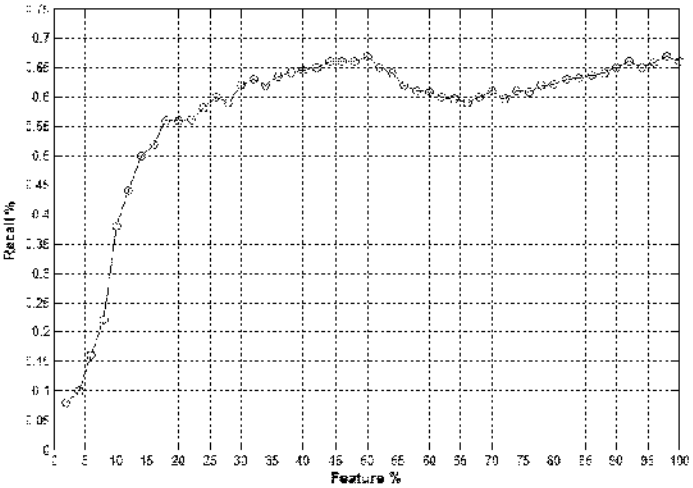

Fig7. Recall rate based on percentage of features

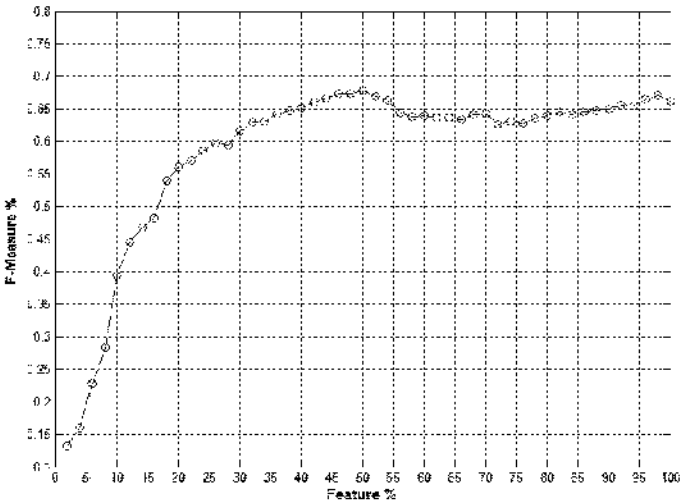

Fig 8.F-Measure based on the percentage of features used

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Figure 6: 50% Features

Figure 7: All Features

REFERENCES

[1] Jia Li, James Z. Wang, ``Automatic linguistic indexing of pictures by a statistical modeling approach, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 9, pp. 1075-1088, 2003.

[2] James Z. Wang, Jia Li, Gio Wiederhold, SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries, IEEE Trans. on Pattern Analysis and Machine Intelligence, (2001) vol 23, no.9, p 947-963.

[3] MULLER, H., MICHOUX, N., BANDON, D., AND GEISSBUHLER, A. 2004. A review of content-based image retrievalSystems in medical applications—Clinical benefits and future directions. *Int. J. Medical Inf. 73*, 1, 1–23.

[4] MULLER, H., MARCHAND-MAILLET, S., AND PUN, T. 2002. The truth about Corel—Evaluation in image retrieval.In*Proceedings of the International Conference on Video Retrieval (CIVR)*. Lecture Notes in ComputerScience, vol. 2383. Springer, 36–45.

[5] ZHANG, H. J., WENYIN, L., AND HU, C. 2000. IFIND—A system for semantics and feature based image retrievalOver Internet. In *Proceedings of the ACM International Conference on Multimedia*.

[6] SMOLKA, B., SZCZEPANSKI, M., LUKAC, R., AND VENETSANOPOULOS, A. N. 2004. Robust color image retrieval forthe World Wide Web.In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

[7] J.F. Silverman, D.B. Cooper, Bayesian clustering for unsupervised estimation of surface and texture models, IEEE Transactions on Pattern Analysis andMachine Intelligence 10 (4) (1988) 482–495.

[8] Noureddine Abbadeni, "Texture representation and retrieval using the causal autoregressive model", J. Vis. Commun. Image R. 21 (2010) 651–664

[9] S.Arivazhagan, L.Ganesan, S.Selvanidhyananthan, " Image Retrieval using Shape Feature", international journal of imaging science and engineering (IJISE),GA,USA,ISSN:1934-9955,VOL.1,NO.3, JULY 2007

[10] Jun Wei Han, Lei Guo, "A shape-based image retrieval method using salient edges", Signal Processing: Image Communication 18 (2003) 141–156

[11] Horn, R. A. and Johnson, C. R. "Norms for Vectors and Matrices." Ch. 5 in *Matrix Analysis*. Cambridge, England: Cambridge University Press, 1990.

[12] I. Kononenko, Estimating attributes: Analysis and extensions of relief, in: Machine Learning:ECML-94, Vol. 784 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 1994,pp. 171–182.

[13] K. Kira and L. A. Rendell, *A practical approach to feature selection*, Proc. 9th Int. Conf. Mach. Learn., (1992), pp. 249 – 256.

[14] M. Dash and H. Liu. Feature selection for classification. *Intell. Data Anal.*1(3): 131–156. 1997.

[15] I. Kononenko. Estimating attributes: Analysis and extensions of Relief. *Proceedings of European Conference on Machine Learning*, Springer-Verlag pp. 171–182. 1994.

[16] I. Guyon, H.-M. Bitter, Z. Ahmed, M. Brown, and J. Heller. Multivariate non-linear feature selection with kernel multiplicative updates and gram-schmidt relief. In *BISC FLINT-CIBI 2003 workshop, Berkeley, Dec. 2003*, 2003

[17] Andras Hajdu, Tamas Toth," Approximating non-metrical Minkowski distances in 2D", Pattern Recognition Letters 29 (2008) 813–821

[18] Jun Ye," Cosine similarity measures for intuitionistic fuzzy sets and their applications", Mathematical and Computer Modelling 53 (2011) 91–97

[19] Y. M. Latha1, B.C. Jinaga and V.S.K. Reddy, "A Precise Content-based Color Image Retrieval: Lifting Scheme", ICGST-GVIP Journal, (2008), 25-32

[20] Tania DiMascio, DanieleFrigioni, LauraTarantino, "VISTO: A new CBIR system for vector images", Information Systems 35 (2010) 709–734

[21]]L. Egghe, "The measures Precision, Recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations", Information Processing and Management 44 (2008) 856–876

[22] Nancy Chinchor, MUC-4 Evaluation Metrics, in Proc. of the Fourth Message Understanding Conference, pp. 22–29, 1992

[23] D. Koller and M. Sahami. Toward optimal feature selection. In *13th International Conference onMachine Learning*, pages 284–292, July 1996.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)