GMM-CR-Based Semantic Image Retrieval and Annotation

Farshad Teimoori, Ali Asghar Beheshti Shirazi,
Electrical Engineering School, Iran University of Science and Technology

***Abstract*— In this article, a content-based image retrieval and annotation architecture is proposed. Its attitude is decreasing the semantic gap. To achieve a narrower gap, the model is based on estimating the relationship between image pixels and image concepts through partitioning the image with unsupervised classifier. Partitioning is executed by dividing the image to its conceptual regions. GMM is the preferred unsupervised classifier and visual features are color and texture of localized windows which sweep the image completely. To decrease the semantic gap, a set of HSV, CIELAB and YCbCr components are used to extract more information as color feature accompanying with dual-tree complex wavelet components as texture feature that can distinguish different patterns more accurately in comparison to other texture extractor. The newly proposed method is evaluated on Corel5K database and its performance is compared with query by visual example and query by semantic example methods comprehensively.**

***Index Terms*—Conceptual regions, Content-based image retrieval, Gaussian mixtures, semantic annotation, semantic retrieval, Query by example.**

I. Introduction

Image databases among different databases have experienced fast growing. Some of most supporting reasons for this so-called explosion in size and number of image databases can be mentioned as follows: huge number of multimedia sources such as armature and well-equipped cameras, a noticeable cut down in price of digital memories and easy-accesses and availability of Internet as an appropriate media to share and extend these databases. These pre-mentioned reasons have left no shortcut to manage, search and explore through these databases, except of developing methods to annotate and retrieve them [1]-[4].

In last decade, magnificent effort has been made that is mostly resulted in useful user-based [5], semi-automatic [6] and even automatic digital image retrieval and annotation systems [7]-[8]. These algorithms have almost established the basic theory needed for developed image exploring systems.

Among two basic approaches, text-based and image-based, there is no doubt that pure text-based approaches cannot catch up with image-based counterparts in performance [4]. Image-based approaches are considered as methods which are based on image features instead of texts. Defining feature as an initial property of an image which is exploited by not complicated procedures and stored in the form of a vector, we can divide image features into two categories: local and general. In this division, features are categorized according to the place where they belong to. If a feature is extracted from a small part inside the image, it is called local feature, while a feature extracted from all pixels of image, is named general feature. Instead of dividing features by location of windows that these features are taken out from, a more suitable division is based on how developed these features are. Regarding this attitude, features extracted for retrieval and annotation systems are divided into low-level and high-level features [9],[10].

By defining features and exploiting them, a latent step is established. This step is located between the raw pixels of image and a concept which exists in the image. On the other word, features are used to estimate the relationship between image pixels and concepts which are inside the image. If a feature cannot extract enough information, or the information which is extracted by the feature needs other features' data to be supplemented or output of feature needs a noticeable amount of post-processing, it cannot help this estimation at the stage, and it is categorized among low-class features. It should be mentioned that low-level features are not necessarily features which are extracted by simple methods [11]. Methods based on low-level features like color, texture and shape; usually apply one feature to construct their feature databases or to compare distinctive images. Their weak performance without focusing on which classification method they use, shows that low-level features cannot exploit enough image information to recognize and estimate the relationship between image pixels and image concepts [12]-[13]. Not only applying one low-level feature will result in low performance, being focused on special feature of image makes the retrieval or annotation system application-dependant. So being concentrated on individual and special feature of image, having no capability of being applied to other databases and weak output of classifiers due to lack of comprehensive data will result in poor performance of retrieval and annotation methods which are mainly based on low-level features[11].

Applying two or more low-level features, not only exploits more necessary image information for using simple comparative classifiers, even based on simple method like Euclidian or Matusita distances, it can also feed more developed classification approaches like Gaussian mixture models or neural networks with enough necessary input vectors to recognize different concepts [4]. In semantic retrieval and annotation attitude, simple low-level features cannot be applied to a progressed classifier like neural network. This impossibility does not illustrate weakness of NNs, whereas it shows that on the one hand NNs needs more comprehensive information and on the other hand they need input vectors that are more distinctive in comparison to other classifier. So, before applying features to NNs, it should be considered that input vectors cover a wide range of samples and input vectors have the maximum variance. To achieve the last goal PCA can be used to maximize the variance of input vectors elements. So, using GMM and NN while regarding these vital details, not applying low-class features directly to GMM or NN and processing the input vectors before applying to these classifiers, can empower a retrieval and annotation method. Using two or more features and processing them can make a retrieval or annotation method to be more close to algorithms that human uses to distinct different images, but it should be considered that applying features which do not suit special application or giving equal or incorrect important factors to low-level features, mostly result in poor performance. This defect, in image processing context, is called "semantic gap" and it has left no way out of poor results of low-level features-based methods, except going toward algorithms based on high-level features [4],[14],[15].

This deficiency, semantic gap, is formed because CBIR systems can not accurately estimate the relationship between image concepts and pixels [16]. In context of semantic gap two points should be mentioned. First, the relationship between concepts and pixels within an image is too complicated. This complexity is originated from difficulty of defining a special concept, or in other words, this complexity is due to difficulty of finding distinctive characteristics of a concept which have the minimum overlap with characteristics of other concepts. For example, if we consider having wheels and glasses when defining 'car' as a concept, we have constructed a new concept that has big overlap with other concepts such as 'airplane'. The second point is lack of exact knowledge about recognition system of human and how brain's neural networks distinguish and memorize a new concept through a huge variety of images which include special concept, have based all the annotation models on an initial and simple sketch of the real annotation system. Needing a more accurate model for annotation system of human has connected this field directly to other sciences and has postponed designing the next generation of annotation systems till having a magnificent progress in other sciences like neurology [17].

II. Related Works

Another attitude toward retrieval and annotation systems, regardless of applying low or high-level features which have been used inside the system, is based on which form of input these systems have utilized. Two principal paradigms have developed over the years: query by visual example (QBVE) and query by semantic example (QBSE)[14]. QBVE is retrieval and annotation architecture that its main input is an image and the whole process of retrieval and annotation is restricted to exploiting and saving low-level features of images (e.g. color histogram) in a depository and accomplishing the retrieval by finding the most similar feature vectors, stored in its database, to the feature vector of an unseen image. In methods mainly based on QBVE [18], comparing low-level feature vectors mean that image similarities are confined to visual similarities. On the other words, in this architecture two images are considered widely similar if they have noticeable similarities in their visual characteristics. As mentioned before, the relationship between image pixels and image concepts is too complex that cannot be estimated by a straightforward approach like QBVE.

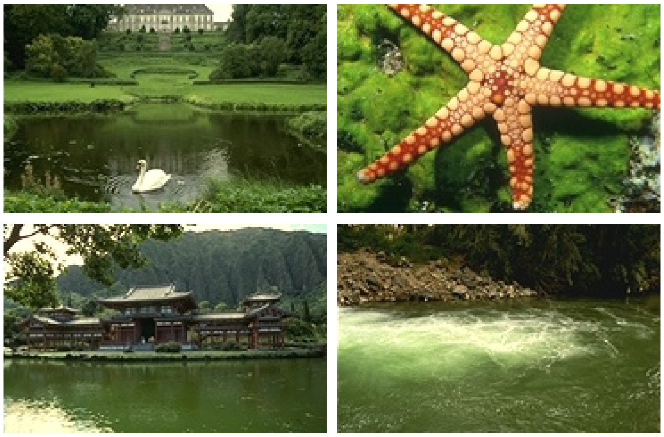

Fig.1. Exceptions of Water as a Concept in Corel5K database

The deficiency of QBVE-based approaches has motivated designing a retrieval and annotation architecture which can simulate human recognition system more accurately. QBSE is a recently developed approach, designed to decrease 'semantic gap'. In this attitude, instead of performing retrieval and annotation algorithms on an image as the input, semantic keywords play the role of system inputs [16]. Comparing QBSE and QBVE, we can consider an elevation in definition of similar images. In QBVE two images are similar if they have an extended similarity in their visual features, but in QBSE attitude, two images are considered similar, if they include similar concepts [14].

Since there is no notion of QBSE semantic architecture in QBVE, they are completely separated paradigms and comparing them on one hand is not simple and straightforward and on the other hand, having different attitude toward similarity, has caused their comparison to some extend far away from being illustrative. If we want to put them in perspective, the first noticeable difference would be generalization. In this context, methods close to the main idea of QBSE outperforms QBVE-based methods. Generalization is the ability of method to find the extreme feature vectors of special concept. In other words, a method has better generalization, if it could find, exploit and make use of all examples of one concept and even its exceptions. As an example, color feature of 'water' as a concept, is mainly blue or white but water of a lake located close to a jungle can reflect the green color of trees. As illustrated in Fig. 1 green is an exception for concept of 'water', 'river' and 'lake'. So, methods mainly based on QBSE attitude, can exploit necessary data of the image which includes an especial concept and even find and use extreme features. Here, extreme features means features of images which are considered exceptions for a special concept.

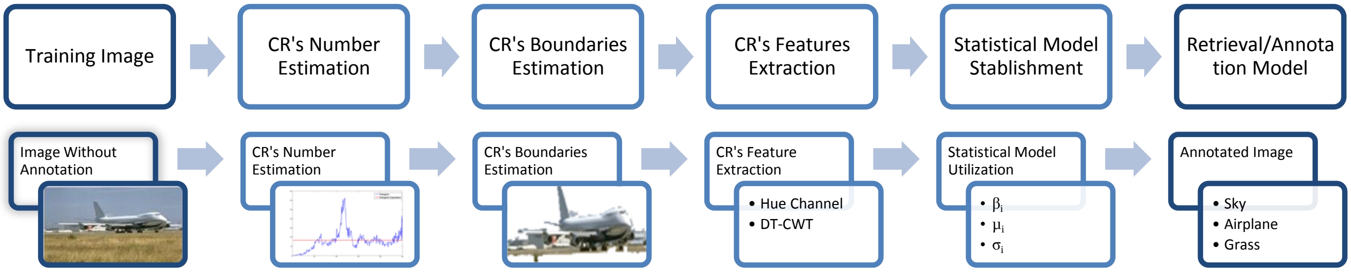

Fig. 2. The main architecture of proposed method: Top procedure shows training and annotation procedure is illustrated in bottom process

Considering better generalization as a virtue of QBSE methods, multiple semantic interpretations has weakened the performance of methods primarily based on semantic model. It means concepts which exist in an image are subjective. This disadvantage is mainly originated from two reasons: first, semantic keywords of QBSE are restricted. In fact a semantic keyword is a concept and a retrieval and annotation system attempts to find the probability of a concept existence in an unseen image. On the one hand the number of concepts cannot be infinite and even cannot be as extended as concepts that human can find in an image and on the other hand, a retrieval and annotation system with a more expanded semantic keywords is not necessarily a preferable system, because a system with a wide range of concepts must apply powerful features to distinguish similar concepts and in this case applying weak features will directly results in poor performance. Second, salient concepts of an image are subjective. In other words the dominant concept in an image is user-dependant and it causes that different users interprets an image differently.

The third difference is related to the keywords space which retrieval or annotation system is trained on. If a an algorithm is trained on a concept which exists in semantic keyword space, that keyword is added to the algorithm space. Apparently, both QBVE and QBSE have shown better results inside their keyword space in comparison to their keywords outspace and it should be mentioned that if QBSE and QBVE are trained on a similar image database, QBSE illustrates better performance. But if their performance in outside of the keywords space is focused, QBVE has shown better performance in comparison to QBSE. It means that a QBVE-based approach can retrieve an image which consists a new concept. A new concept is a concept that the method is not trained on. This breakdown is caused by the form of query which applied to a QBSE-based method, i.e. in QBSE all the information of image is compressed in limited number of words while the query of a QBVE-based method is a complete image that even includes the concept which the method is not trained on.

In last paragraphs, privileges and deficiencies of QBVE and QBSE are mentioned. In proceeding paragraphs, both approaches would be considered mathematically. An image database, $\mathrm{D} = \{\mathrm{I}_1, \dots, \mathrm{I}_\mathrm{N}\}$, is the initial point of any retrieval and annotation system. N is the number of image in the database and I_i stands for ith image. Each system has its special feature space χ, that images are considered to be its observations. In the case that database images have no labels, each image is an example of a different class, illustrated by a random variable Y, defined on $\{1, \dots, \mathrm{N}\}$. Architectures based on this attitude is said to function at the visual-level and in minimum probability of error sense, given a query image I_q, it must be assigned to the largest posterior probability, i.e.,

$$y^* = \arg\max_y P_{y|x}(y \mid I_q) \tag{1}$$

In semantic attitude, there is a keyword space that plays the rule of Y. Here, instead of assigning a number to an image, each image is given a vector C_i. Supposing semantic keyword space $\varpi = \{w_1, \dots, w_L\}$, this vector would be an L-dimensional vector that its elements $c_{i,j}$ is 1 if the ith image is annotated with jth keyword. These words play the rule of classes or in other words they form the concept space. In practical systems, since the number of keywords that can be attached to an image is restricted, there are concepts existing in the image but not attached to it. So image databases usually are weakly annotated and images are labeled with concepts which are seemed more relevant to the labeler [14]. In this article we suppose that image database is weakly labeled.

In general, a semantic retrieval and annotation system has two main steps: establishing an statistical model that estimate the relationship between image concepts and image pixels and applying this model for an unseen image. In this attitude, concepts are words which shape different classes and an image is annotated when the most probable concepts of that image is connected to it as a form of one or more words.

At semantic level (a QBSE-based approach) a random variable W is defined which takes values in $\{1, \dots, L\}$. If each image has a set of n feature vectors, $I = \{\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{n}\}$, where $\mathrm{x}_\mathrm{i} \,\epsilon\, \chi$, $W = i$ is held if and only if x is a sample of the concept w_i. Given a new image I, MPE annotation must label it with concept of largest probability

$$w^* = \arg\max_w P_{w|x}(w \mid I) \tag{2}$$

Accomplishing the annotation of all unseen images in database, it is possible to start retrieval. Supposing w_q a query concept, the MPE sense retrieval's function is to select images with largest posterior annotation probability.

$$i^* = \arg\max_i P_{x|w}(I_i \mid w_q) \tag{3}$$

III. Proposed Query by Semantic Example Annotation System

Annotation system proposed here utilizes high level features. Its input is an unseen image and the output is a 1×L vector that its ith element is 1 if w_i concept exists in the image. The system is based on MPE to annotate an image and it includes two major procedures: training and annotation. The detailed processes are depicted in fig. 2. In this figure the first row shows training procedure and the second one illustrates different steps in annotation.

In training procedure (upper row), the input is a labeled image. At first step, number of conceptual regions (CR) of the image is estimated through its histogram. In next step boundaries of CRs are approximated. These boundaries are determined by usage of image visual features. Then the image is partitioned to its conceptual regions as these regions have no overlap and they cover the whole image. Next, visual features of CRs are computed and saved as the image features in the form of a matrix f_j. These processes are done for all training images which consist an special concept and eventually data associated with the concept is available in a matrix F_i. The final step in training procedure is establishing an statistical model for each concept based on $\underline{F_i}$. This model, named π_i, estimates the relationship between a concept and its visual features matrix.

$$w_i \Leftrightarrow F_i \Leftrightarrow \pi_i \tag{4}$$

Some semantic retrieval or annotation methods have mentioned a variable, located between the input and the final results of their statistical model. This variable is mostly named latent variable and in the model proposed in this article, visual features play the same role as latent variable. In training procedure which is the most time consuming step, the input is a huge number of training images and the final outcome is an annotation model for each concept in the concept space.

After accomplishing the training, for each concept, w_i, all parameters of statistical models, π_i, are available for annotation procedure. Similar to training steps, the number and boundaries of CRs (2nd and 3rd steps in fig.2-1.b) must be estimated for an unseen image. Here an unseen image is an image that is not used in training step. After partitioning the image is done, visual features must be extracted from CRs. Features extracted in this step are exactly from the same type that have been exploited in training. Eventually, these visual features are applied to statistical model, π_i, prepared from training, and annotations of image are obtained.

In proposed annotation architecture, our tendency for estimating the relationship between image pixels and image concepts is based on approximating the number of CRs, their associated boundaries, computing CRs' visual features and establishing an statistical model in the form of GMM's parameters in two main procedures.

IV. Approximating Number of CRs

An image is a set of pixels located next to each other and is generally saved in a digital form of three values per pixel. The most valuable and easy-extracting information of image exists in color data. In our process, a conceptual region of image is a part of image that has four properties:

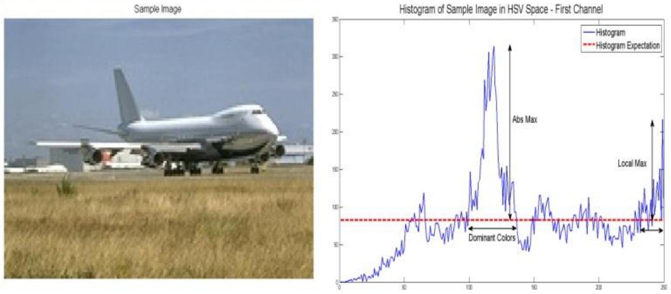

Fig. 3. Sample Image and Its Hue Hisogram

- *A CR is a united and continuous part of image*
- *There is no hole inside a CR*
- *CRs have no overlap with each other*
- *All CRs build the whole image*

The mathematical definition of CR mentioned above, on the one hand prepare a simple definition which can be executed with noticeably low computation cost and on the other hand, it partitions the image in a way that is so close to object partitioning methods. In partitioning method used in this article, at first step, the image is transformed into standard HSV color space. The advantage of HSV color space is that it constructs the most similar color interpretation to what human realizes from color [19]. So partitioning the image into its CRs in HSV space causes that conceptual parts of image evoke the parts distinguished by human's brain. At next step, a 512-bin histogram of hue channel of image is obtained. This data is needed to estimate the number of CRs.

Before using data obtained in last step, we need a new definition. Histogram Expectation (*HE*) is an statistical definition and is originated from considering the whole image as a CR. HE is defined the division of number of image's pixels by number of histogram's bins. Apparently, HE is equivalent of a monotonous image. Another definition which is needed to estimate the number of CRs is Dominant Color (*DC*). DC is a bin of image histogram which has a value more than image's *HE*. To have executive formulas for new definitions, image width and height are shown with w and h respectively, number of histogram's bin is denoted with N_h and n_i is used for a histogram's bin. *HE* and *DC* are obtained:

$$Histogram\ Expectation\ (HE) = \frac{w \times h}{N_h} \tag{5}$$

$$Dominant\ Color\ (DC) = \{n_i \mid n_i \geq HE\} \tag{6}$$

*DC*s affect the number of CRs directly. It means if the number of dominant colors in an image increase, more number of CRs is needed to cover the whole image conceptually. Fig.4-1 shows the histogram's characteristics which are involved in estimation of number of CRs. Besides *DC*s, absolute maximum of histogram is another factor which affects the number of CRs. Bigger absolute maximum in the image histogram means the biggest CR in the image covers bigger region of image. So, it affects the number of CRs inversely and greater absolute maximum means less complicated image and then fewer number of CRs is needed. The last factor which is involved in this estimation is number of local maximums. Unlike absolute one, it affects the estimation directly. An image which has more local maximums needs more CRs to be covered conceptually. Their effects in estimating the number of CRs is so similar to *DC*s' effects.

V. VISUAL FEATURES

Among different visual features, color has a special ranking and properties. On the one hand it consists huge portion of information that exists in an image and on the other hand, in comparison to other features, extracting and exploiting its information are more straightforward. In annotation method proposed in this article, for compensating the weakness of using one color space, three color spaces are employed. HSV, YCbCr and CIELAB are spaces that are employed. As mentioned before, HSV prepares the closest interpretation of what human being recognizes from color [19]. YCbCr is preferred over the other spaces because from information theory point of view, most of color information exists in its first two channels. In CIELAB color space, color's shadows are omitted in its first two channels [4]. In CR's boundaries estimation four color channels are used: first channel of HSV, second and thirds channel of YCbCr and first channel of CIELAB. Although four color channels are utilized in boundaries estimation, just HSV's first channel is employed as a color feature of CRs.

Considering the color as the most effective feature in annotation and retrieval algorithms, texture can prepare alternative information for completing data that is extracted by color features [4]. Among different methods of texture data extraction, co-occurrence matrix, Tamura features, Wold model, discrete feature transform, real wavelet transform and complex wavelet transform can be mentioned. First Three texture extracting methods are not employed in the proposed method due to their heavy computational cost and obvious deficiency [20]. DFT can only distinguish texture patterns which are frequently repeated. Unsatisfactory sensitivity to shift, poor performance in finding the angle of a pattern and being improper for analyzing high frequency signals with narrow bandwidth demonstrate disadvantages of DWT and decrease texture features' compare to dual-tree complex wavelet transform [20]. In the proposed method, 2-D CWT are employed which decomposes a matrix f(x,y) using dilation and translations of a complex scaling function and results in six complex wavelet functions:

$$f(x,y) = \sum_{k\epsilon z^2} S_{j_0,k}\phi_{j_0,k}(x,y) + \sum_{b\epsilon\theta}\sum_{j\geq j_0}\sum_{k\epsilon z^2} c^{\theta}_{j,k}\psi^{\theta}_{j,k}(x,y) \qquad (7)$$

$$\phi_{j_0} = \phi^{l}_{j_0} + \sqrt{-1}\phi^{i}_{j_0} \qquad (8)$$

$$\psi_{j_0} = \psi^{l}_{j_0} + \sqrt{-1}\psi^{i}_{j_0} \qquad (9)$$

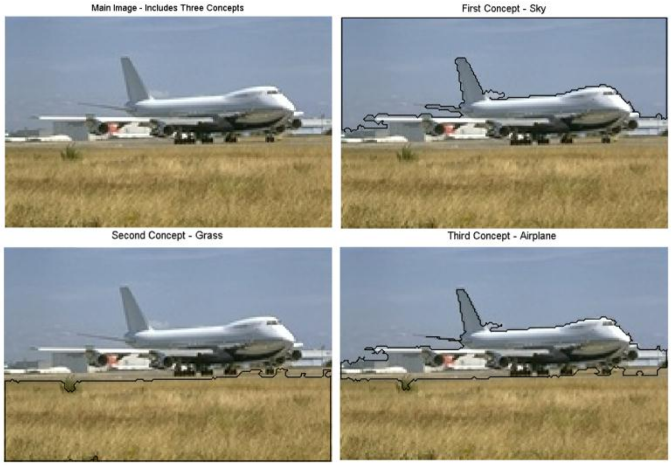

Fig. 4. Sample Image Includes 3 Concepts and Its 3 CRs

$\psi(x)$ is the mother wavelet and $\phi(x)$ denotes scaling function whereas $S_{j_0,l}$ is scaling coefficient and $c_{j_0,l}$ is complex wavelet coefficient . Actually CWT is combination of two real wavelets and it inherits the computational efficiency of separable transforms. In proposed method, a 2-D CWT with three steps are used. This transform can distinguish $\pm 15, \pm 45$ *and* ± 75 besides 0, 360 and ± 90 directions. So, besides a powerful color feature described in this section, the employed texture feature is approximately robust to all texture directions [20].

VI. CR'S BOUNDARIES ESTIMATION

After estimating the number of CR's, in order to find CR's boundaries, a 9 ×9 pixels window which has 3 pixels overlap with its neighbors covers the image in horizontal and vertical manner. For each 9×9 window, visual features, as explained in section V, must be computed. For each local window, data in four color channels is saved in a vector x_i and all these vectors form a matrix x which consists all visual features of the image. In this stage, because there is no training data, an unsupervised classifier is needed. Considering GMM's performance, its noticeable speed in reaching the final parameters and fast two steps categorizing algorithm, GMM is exploited in the estimation architecture proposed in this article [15]. In fact, GMM is a linear weighted combination of Gaussian distributions [14]:

$$P_{X|L}(x,l) = \sum_{i=1}^{n} \beta_i \, \mathcal{G}(x, \mu_i, \sigma_i) \qquad (10)$$

In this relation, n ,the number of CRs, is the only parameter which its value is calculated before. $\mathcal{G}$ denotes Gaussian distribution and x is a matrix containing image visual features. In fact, x is a l_w×324 matrix, while l_w is the number of windows needed to cover the image and data extracted from four color channels are saved in 324 elements. In other words, for each 9×9 window, there are 324 samples. In relation 10, Gaussian distribution's weights β_i, their averages μ^l_i and their variances σ^l_i are not determined. Considering their initial values and running the iterative algorithm of expectation maximization (EM), their final values would be reached. After executing the GMM classifier on matrix x, parameters of n

Gaussian distribution are determined. Then, by applying each rows of x to n Gaussian distribution, n related probabilities are obtained and the maximum probability indicates the class which the local window belongs to. Eventually each local window belongs to one category and estimation of CR's boundary are done.

Fig. 4 shows the final result of executing the CR's boundary estimation on a sample image. Main picture and three concepts exists in the image are depicted. Image conceptual regions are 'sky', 'grass' and 'airplane' respectively. In the partitioning method applied here, non-overlapping principal is regarded: CR's have no overlap and they build the whole image together. In this stage, CR's number is the input and the output is CR's boundaries.

VII. Statistical Model

After estimating the number of CRs and their boundaries, the image is partitioned to its conceptual parts. Next step is extracting visual features of these CRs. Each region is considered individually and both color and texture features are calculated and are saved in a CR_{i}×164 matrix, $\mathbb{X}$, which CR_i is the number of CRs of ith image. The first 128 elements of $\mathbb{X}$'s row belong to color feature and the left 36 elements belong to texture feature.

In training step, all the images which have a special concept, for instance 'sky', are found. Then, their final matrixes, after estimating the number of CR, their boundaries, and extracting CRs' visual features are calculated and are stored in matrix $\mathbb{X}_j$. These matrixes form the feature matrix of the concept $\mathbb{C}_i$. Considering that each training images include more than one concept, the matrix $\mathbb{C}_i$ includes features of more than one concept. Since in establishing $\mathbb{C}_i$, a special concept has been concentrated, we expect that ith concept, for instance 'sky', owns the most samples.

After computing $\mathbb{C}_i$, since there is no training data, in order to classify the concepts and find the frequent properties of the expected concept, an unsupervised classifier is needed. Based on the reasons mentioned before, GMM is preferred and is executed on the $\mathbb{C}_i$. The output of GMM algorithm is π_i, the statistical model for the ith concept. As an example, executing the GMM on the sky's matrix $\mathbb{C}_{sky}$, would result in its conceptual model π_{sky}. By conceptual model, we mean Guassian's parameters: its weights, averages and variances. If w_i is a concept in keywords space, a conceptual model is associated to each w_i:

$$\{" w_i \hat{I} \; v \} \hat{U} \; \{\mathrm{C}_{\mathrm{i}}\} \hat{U} \; \{b_i, m_i, s_i\} \tag{11}$$

After accomplishing the training and finding all $\mathbb{C}_i$, the input of annotation procedure is an unlabeled and unseen image. This image consists some concepts and after annotation, these concepts would be attached to the image in the form of some words. At first, similar to training step, number of CRs is estimated. Then their boundaries are approximated and in the next step visual features of there CRs are computed. Eventually, vector of CR's visual feature are applied to all $\mathbb{C}_i$ and for each concepts there would be a probability. Actually these probabilities show the probability of concepts' existence in the image. After ordering these values in descending manner, concept associated with first five probabilities are attached to image as its annotations.

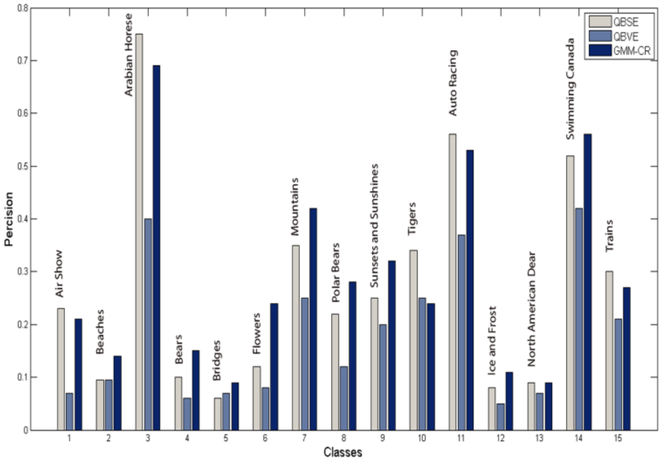

Fig. 5. MAP scores of GMM-CR vs QBSE and QBVE

VIII. Proposed Method Evaluation

A. Image Database

In order to evaluate the annotation and retrieval algorithms, an image database is needed and in order to compare the performance of different algorithms, different databases are created and managed by experts. Performance evaluation of a method is meaningful if it is executed on a relevant database.

On the one hand, number of images in a database must be numerous enough that strong and weak points of method under evaluation can be determined and on the other hand the image database must be comprehensive. In other words, image database must consist enough number of photos and concepts.

Considering these, Corel5K is selected to evaluate the proposed method. This database includes 5000 images that 4500 and 500 images are selected for training and assessment respectively. Images in these two categories are selected properly and in both training and annotation steps, there are enough images for each concept. Corel5K includes 371 concepts and each image is annotated by 5 words. Huge number of images and concepts in Corel5K shows its comprehensiveness and it is bound with massive amount of calculation in training and evaluation stages.

B. Exprimental Evaluation

To evaluate different methods of retrieval and annotation, precision and recall are classical scales for performance evaluation. They are widely used and adopted by TRECVID assessment benchmark [14]. If the number of retrieved images which are relevant is shown by "r" and the number of relevant images in the image database is denoted by "R", precision is "r/N" while N top database matches are chosen. Considering these parameters, recall is defined by "r/R".

Fig. 5 compares performance of QBVE, QBSE and GMM-CR. Here QBVE and QBSE mean a QBVE-based and QBSE-based method respectively which are evaluated in [14]. As it can be seen, QBVE overcomes QBSE just in one category, 'Bridges', and QBVE has a better performance versus GMM-CR just in 'Tigers' class. As mentioned before QBSE/QBVE

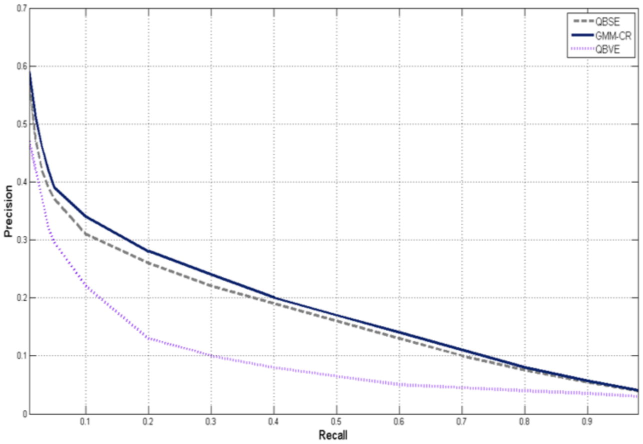

Fig. 6. Precision-Recall diagrams of GMM-CR vs QBSE and QBVE

comparison cannot be illustrative noticeably, while QBSE/GMM-CR comparison can elaborate why GMM-CR has overcome QBSE in 9 classes out of 15 classes.

As Fig. 5. shows, in 'Beaches', 'Bears', 'Bridges', 'Flowers', 'Mountains', 'Polar Bears', 'Sunsets and 'Sunshine', 'Ice and Frost' and 'Swimming Canada' classes, GMM-CR outdoes QBSE. Comparing the concept that GMM-CR overcomes QBSE in, shows that CRs associated with these concept, in comparison to the concept that proposed retrieval and annotation method is defeated by QBSE, covers a bigger regions of the image. In other words, GMM-CR has a better performance in classes which their related CRs occupy a bigger part of the image. For instance in 'Ice and Frost' and 'swimming' classes, most of images are covered with white white and white/blue colors respectively. So, partitioning the image to its CRs and finding the visual features of these CRs feed the proposed method with enough data to be well trained in these classes whereas this data might be not comprehensive in classes like 'Tigers'. Another point that can be mentioned about fig. 5. is that QBSE defeats QBVE with obvious difference, but GMM-CR overcomes QBSE slightly.

Another meaningful comparison is contrasting the precision-recall diagrams, as it is depicted in Fig. 6. The apparent improvement of QBSE and GMM-CR versus QBVE is due to elevation of similarity definition [14]. As mentioned before, in QBVE attitude two images are similar if they are comparable in visual features space, but in QBSE and GMM-CR architectures, two images are considered widely similar if they both include same concepts. This improvement has made these two architectures closer to retrieval and annotation model used by human's brain. But the slight improvement in GMM-CR versus QBSE is due to smaller semantic gap. This improvement is originated from definition of conceptual region and its success in simulating the realistic model of retrieval and annotation.

IX. Conclusion and Future Work

In this work, we have presented a new CBIR retrieval and annotation model which is mainly focused on decreasing the semantic gap. Our attitude toward a narrower gap has been based on estimating the relation between image pixels and concepts more accurately. Partitioning the image to it conceptual regions based on estimating the number of CRs and their boundaries, extracting visual features, mainly color and texture, and establishing and statistical model in form of GMM's parameters are executed to find the characteristic of each concept comprehensively. Although we tried to use more accurate model to estimate the relationship between image pixels and concepts, we have used four color channels data to extract more color information and a more powerful texture feature are used to make the color information more complete to feed the GMM as the classifier.

The model is evaluated on 5000 images (Corel5K) and its performance is compared to QBSE and QBVE. GMM-CR model overcomes QBVE with noticeable difference as well as QBSE does, but it beats QBSE slightly. Since GMM-CR is successful in retrieving and annotating concepts which they cover a continuous part of the image, it has a slightly better performance in comparison to QBSE. For future works, this weakness can be compensated and GMM-CR's performance in images with large number of CRs can be improved.

References

[1] Luo, J. B., Boutell, M., & Brown, C. (2006). "Pictures are not taken in a vacuum: An overview of exploiting context for semantic scene content understanding", *IEEE Signal Processing Magazine*, 23(2), 101–114

[2] Y. Rui and T. S. Huang, "Image Retrieval: Current Techniques, Promising Directions, and Open Issues", *J. Visual Communication and Image Representation*, Vol.10, pp.39-62, 1999

[3] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Reading, MA: Addison-Wesley,June, 1999.

[4] Yu-Jin Zhang, *Semantic-based visual information retrieval,* IRM Press, 2007.

[5] Lieberman, H., Rosenzweig, E., & Singh, P. (2001). "Aria: An agent for annotating and retrieving images". *IEEE Computer*, *34*(7), 57–61

[6] R. Srihari, & Z. Zhang, "*Show&Tell: A semi-automated image annotation system*". IEEE Multimedia, 7(3), 2000, 61–71

[7] J. Li, J. Wang, "Automatic linguistic indexing of pictures by tatistical modeling Approach". IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(9), 2003, 1075–1088

[8] N. Vasconcelos and M. Kunt, "*Content-based retrieval from image databases: Current solutions and future directions,"* in *Proc. Int. Conf. Image Processing*, Thessaloniki, Greece, 2001.

[9] S. Zhu and Y. Liu,"Scene segmentation and semantic representation for high-level retrieval", *IEEE Signal Processing Letters*, Vol. 15, pp.713-716, 2008

[10] R. Picard, "Digital Libraries: Meeting Place for High-Level and Low-Level Vision". Paper presented at the Asian Conference of Computer Vision, Singapore, 1995

[11] J. Yu and Q. Tian, "Semantic subspace projection and its application in image retrieval", *IEEE Transactions on Cicuits and Systems for Video Technology*, 2008, 18, (4), pp. 544-548

[12] A. Jain and A.Vailaya, "Image retrieval using color and shape", *Pattern Recognit. J.*, vol. 29, Aug. 1996.

[13] R. Manmatha and S. Ravela, "Asyntatic characterization of appearance and its application to image retrieval," *Proc. SPIE*, vol. 3016, 1997.

[14] N. Rasiwasia, P J Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example", *IEEE Transactions on Multimedia*, 2007, 9, (5)

[15] G. Carneiro, Pedro J Moreno, and Antoni B Chan, "Supervised learning of semantic classes for image annotation and retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29, (3)

[16] Gustavo Carneiro, Antoni B. Chan, Pedro J. Moreno, and Nuno Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 3, March 2007

[17] Simon Haykin, Neural Networks- A Comprehensive Foundation, Prentice Hall International , 1994

[18] W. Niblack, "The QBIC Project: Querying Images by Content Using Color, Texture, and Shape". Paper Presented at the Conference of

Storage and Retrieval for Image and Video Databases, San Jose, CA, USA, 1993

[19] R. C. Gonzalez and R. E. Woods, Digital Image Processing, Prentice Hall, 2002, 2nd

[20] M. Kokare, P. K. Biswas, and B. N. Chatterji, "Rotation-Invariant texture image retrieval using rotated complex wavelet filters", *IEEE Transactions on Systems*, Man and Cybernetics, 2006, 36, (6)

[21] A. Jain and A.Vailaya, "Image retrieval using color and shape", *Journal of Pattern Recognition*, 1996, 29

Farshad Teimoori Received his B.S. in Electrical Engineering (Communication) in 2007 from Shahed University, Tehran, Iran and accomplished his M.S. in the same field in 2010 in Iran University of Science and Technology (IUST). He started his Ph.D course in Islamic Azad University, Tehran Science and Research branch in 2011 in Telecommunication Engineering. His research interests are Multimedia Processing especially Image Processing (Image Annotation, Image Retrieval, Image Registration), Speech Processing (Speech Enhancement, Speech Recognition, Speech Coding) and Digital Signal Processing.

Ali Asghar Beheshti Shirazi received his B.S. and M.S degree in Communication Engineering from Iran University of Science and Technology (IUST) in 1984 and 1987 respectively and Ph.D from Okayama University, Japan in 1995. He joined the School of Electrical Engineering, IUST where he is currently Assistant Professor. His research interests include Digital Image Processing, Data Communication Networking and Secure Communication.