# ORAL-QUERY-BY-SKETCH: AN XML-BASED FRAMEWORK FOR SPEECH ACCESS TO IMAGE DATABASES

Shiow-yang Wu
*Department of Computer Science and Information Engineering*
*National Dong Hwa University*
*Hualien, Taiwan, R. O. C.*
showyang@csie.ndhu.edu.tw

Wen-Shen Chen
*Department of Computer Science*
*and Communication Engineering*
*Dahan Institute of Technology*
*Hualien, Taiwan, R. O. C.*
tedc@ms01.dahan.edu.tw

**Abstract** We propose an XML-based framework for speech access to multiple image databases without the need for keyboard or pen-based input. The key idea is to orally command the system in drawing an abstract sketch of the target images using simple graphical objects. We define an oral language for image description and query control named SpeechQuel. The abstract sketch is represented in W3C SVG format while all other query constraints are represented using a device and system independent XML language called SpeechQuelX. This facilitates content-based access to multiple and/or heterogeneous image databases with a single query. Preliminary implementation result successfully demonstrates the feasibility and efficiency of our approach.

## 1. INTRODUCTION

Among all the potential applications of multimedia access, many situations demand easy and hand-free operation to query information from heterogeneous data sources. In such cases, speech accessing methods are the most convenient and promising approach. For disable users, speech access may be the only acceptable method. We propose an XML-based framework (entitled Oral-Query-by-Sketch) for integrated speech access to multiple image databases without

the need for keyboard or pen-based input. The key idea is to orally construct an abstract sketch of the target image(s) using simple graphical objects. An image description and query control oral language called SpeechQuel is defined to command the system in drawing such an abstract sample for querying the image databases by content. To facilitate unified access to multiple and/or heterogeneous image databases, all query constraints are represented using a simple and system independent XML language named SpeechQuelX while the abstract sketch in W3C Scalable Vector Graphics (SVG) format respectively. A new image database can be added into the access range with a simple SpeechQuelX/SVG to native query format translator. This can be done in a straightforward manner with the XSLT technology and any standard SVG to acceptable graphic format convertor. Preliminary implementation result successfully demonstrates the feasibility and efficiency of our approach. As a summary, the main contributions of this paper are:

- The Oral-Query-by-Sketch framework for accessing image data sources based on orally constructed abstract sketches.
- An XML-based approach for unified access to multiple and/or heterogeneous image databases.
- An oral language SpeechQuel for sketch description and query control.
- A device and system independent XML language SpeechQuelX to represent the image queries.
- The application of XSLT technology for translating SpeechQuelX queries into native image database query constraints.
- Several useful criteria for empirically evaluating the effectiveness of query by example image retrieval methods.
- Techniques for integrating off-the-shelf speech recognition interfaces with multiple image databases from different vendors.

The rest of the paper is organized as follows. Section 2 provides a background survey of related issues and research work. Section 3 presents our framework and system architecture. Section 4 describes the SpeechQuel language and the associated SpeechQuelX representation format. In Section 5, we report a preliminary implementation of our framework and propose several useful criteria for empirical evaluation and performance comparison of related content-based image retrieval methods. The evaluation results demonstrate the feasibility and performance of our framework. Section 6 concludes the paper.

## 2. RELATED WORK

The development of image retrieval techniques can be traced back to 1970's where the main idea was to employ text-based retrieval on human annotation of images [2]. With major difficulties such as vast amount of labor required as well as the annotation impreciseness due to human perception subjectivity, these techniques are still useful for processing text-based information associated with images. Content-based image retrieval techniques emerge since the early 1990's [1, 9]. These techniques explore content-based features such as color [12], shape[12], and texture[7] for more accurate and efficient access. Since then, a number of content-based image retrieval systems such as QBIC[4] and MARS[5] have been developed and their performance reported. Among all the previous work, only a limited number of projects were reported on speech based access to document database[11], let alone image data sources. There were a number of systems along the line of query-by-sketch [10, 8, 13]. However, most (if not all) of them were based on hand-drawn sketches. The Oral-Query-by-Sketch approach reported here is unique in that we provide a framework for speech access to image sources based on orally described abstract sketches composed of elementary graphical objects. This is not only useful for normal usage but also well suited for situations that demand hand-free access such as in mobile environments or for disable users. Furthermore, we employed XML technologies[3] and the SVG format[6] to facilitate unified access to multiple image databases using the same user interface and query. A special effort was also placed on the smooth integration of off-the-shelf speech interfaces and image retrieval systems from different vendors.

## 3. SYSTEM FRAMEWORK AND ARCHITECTURE

The Oral-Query-by-Sketch framework and system architecture were designed with several criteria in mind. First of all, we want to enable hand-free and speech access to multiple(possibly heterogeneous) image sources. This calls for a simple yet effective oral language for image queries and the accompanied system architecture for integrating the speech interface with different sources. The goal is especially challenging for the nature of the targets. On text based information, there are many speech to text systems available for spoken language access. For image information, however, content-based retrieval techniques are much more desirable. It is unclear how to seamlessly integrate speech interface into such techniques for even a single database, let alone multiple or even heterogeneous sources. Furthermore, many existing content-based retrieval methods require several rounds of image selection and feedback process to reach the final target. This is in general not desirable or even unacceptable in the intended situations such as mobile environments or for disable users. There is clearly a need for far more simple yet precise content-based access methods.

Our solution to these challenges is the framework and architecture as depicted in Figure 1, a deliberately designed oral language SpeechQuel for image query, as well as an XML-based representation method to cope with heterogeneity. The key ideas are as follows:

- The user employs the image description and query control language SpeechQuel to orally command the system in drawing an abstract sketch of the target image(s) using simple graphical objects.
- The speech input is processed by an off-the-shelf speech recognition software and converted into text-based SpeechQuel commands.
- The system follows the commands in drawing and displaying the abstract sample image. User can construct the sample from a template library or from scratch. User can also refine the sample interactively with visual properties editing commands. Query constraints and nonvisual properties can be specified easily as well. Details of the SpeechQuel language will be discussed in the next section.
- The internal representation of the sketch is divided into two parts. The image itself is represented using the standard SVG format. All other information are represented using an XML language named SpeechQuelX. These include all other query conditions and nonvisual property settings.
- For each image information source, such as an image database, a corresponding query generator is needed to convert the SpeechQuelX/SVG representation of the sketch image and query conditions into native query format for the retrieval of images that are similar to the sample. With the use of XML and standard SVG representations, the query generator can be easily constructed using XSLT and standard SVG conversion technologies.
- Query results from different sources are combined and send back to the user. If the results contain the target image(s), the process ends with a success. Otherwise the user can provide feedback and/or refinement of the sample to start another round of image retrieval.

Orally constructing an abstract sketch using simple graphical objects for content-based access is probably the most distinctive feature of our framework. The system architecture is also designed to be highly modular in the sense that the speech interface as well as the image databases can be basically any exiting and/or off-the-shelf packages or systems. A new image source can be added into the access range as long as it supports a standard SVG to acceptable graphic format convertor and a simple SpeechQuelX to native query format translator. Since SVG is a widely supported W3C format, and SpeechQuelX is a simple

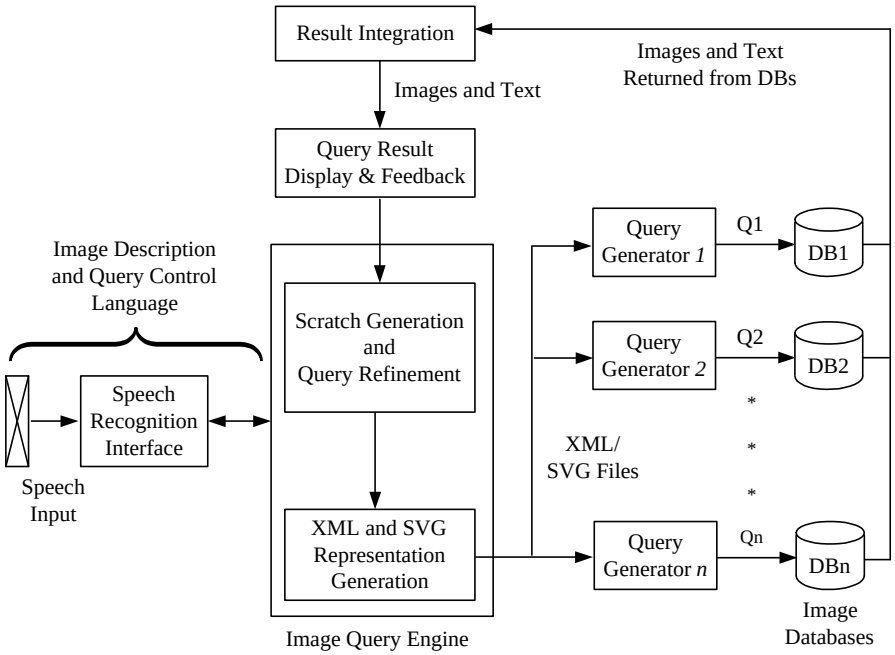


*Figure 1.* The Oral-Query-by-Sketch system architecture.

XML language, building such a convertor or translator is a rather straightforward task. In fact, we provided a simple function in the prototype system to convert an SVG file into JPEG format, and employed the XSLT technology to specify the automatic translators for all the image databases we tested without any difficulty. The key benefit of using the SpeechQuelX/SVG intermediate format is to cope with the potential heterogeneity of different image sources. It also facilitates independent technology and/or system improvement of individual module without affecting the others.

Another key characteristic of our framework is that, by using abstract sample image as the starting point, all image databases we tested were able to quickly retrieve the target images within one or two rounds. This significantly reduced the number of images required to be transmitted between the user and the image servers in comparison with the normal retrieval methods. This characteristic not only reduces the response time significantly but also makes our techniques well suited for bandwidth limited access such as in mobile environments.

## 4. THE SPEECHQUEL LANGUAGE

Drawing a desired image with oral commands is far from easy. We were first inspired by the technique of hand drawing crime suspect based on witness

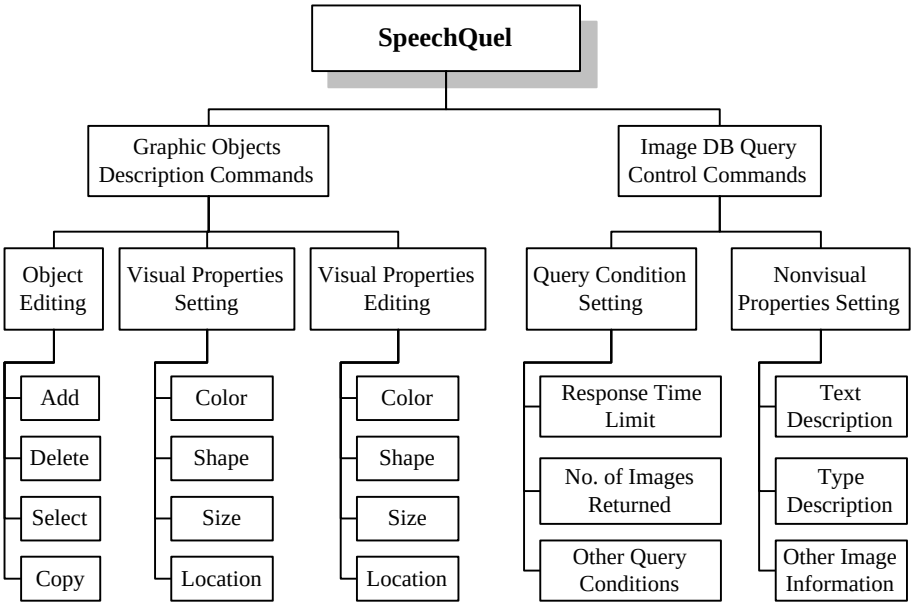


*Figure 2.* The SpeechQuel Language.

description. The pursue of image details is soon abandoned since it is quickly becoming clear that there is no way to define an oral command language that can cover even the modest level of details. Even with such a language, the user is probably unwilling to learn and remember all the commands for describing image details. Besides, describing such an image may take a considerably long period of time. The key point is that our purpose is to quickly retrieve the target image(s), rather than coming up with a fine drawing. For content-based retrieval, we may not need all the details. Therefore we came up with the idea of composing an abstract sketch using simple graphical objects such as circles, triangles, squares, etc. The whole idea is materialized in the language SpeechQuel(Figure 2).

The SpeechQuel language commands are divided into two categories:

- Graphical objects description commands: These are commands for drawing and editing the sample image based on the composition of elementary graphical objects such as circles, squares, triangles, etc.
- Image database query control commands: These are commands for setting the query constraints such as the response time limit, as well as specifying other nonvisual properties.

An important design consideration here is to come up with a natural and minimal set of oral commands that are versatile enough to compose the sample

image user desired. We provide graphical object creation and editing commands to add, select, delete, copy, remove objects. Visual properties such as color, shape, size, and location can be specified or modified with corresponding properties setting/editting commands. There are also commands for moving selected objects, enlarging or reducing object size, as well as saving completed sample images as templates.

To allow finer control over the image retrieval process, another set of commands is provided for setting query constraints and nonvisual properties. The user can specify the response time limit to get whatever result the system retrieved so far. The number of images transmitted between the DBs and the user can be specified to reduce the bandwidth requirement. Users are also allowed to enter text-based information which can sometimes provide indispensable hints in identifying the target images.

We reemphasize the important idea of specifying an abstract sample instead of a detail image. As will be presented in Section 5, with proper shape and color, a simple image can be very efficient in helping the underlying database systems quickly retrieve the target image(s).

Once the abstract sketch is completed, it is represented using an XML/SVG format. The image itself is coded in W3C SVG format for better interoperability. All other information are represented with the XML language SpeechQuelX. The language is designed with tags for most of the commonly used query control commands/settings of existing content-based image databases. We also add some less commonly used query constraints for mobile users such as the response time limit and the maximum number of images allowed to be transmitted between the databases and the user.

The actual retrieval of the images is facilitated by a query generator which simply converts the SpeechQuelX/SVG representation into native query format. A new query generator must be constructed for adding a new image database with different query format into the access range. As stated earlier, however, constructing such a translator is a rather straightforward task with the help of standard XSLT and SVG technologies.

We demonstrate the entire process by going through a sample session in retrieving an image of two kitties. The target and the abstract sketch are shown in Figure 3. The abstract sketch is constructed using SpeechQuel commands as follows (originally in Mandarin and translated into English).

1 "Background color"_"Green"

2 "Add Object"

3 "Shape"_"Ellipse", "Color"_"Brown", "Size"_"Medium",
"Location"_"Center"

4 "Copy Object"_"One"


*Figure 3.* A target image and the abstract sketch for its retrieval.

Note that when a command such as "Copy Object" is issued, the system will number all available objects for selecting the target. The "one" in the last command denotes the selection of the first object on the list for copying.

With the just constructed abstract sketch, we can further instruct the system to search a target database and return a specified number, say 5, of closely matched images. The query conditions and the abstract sketch are transformed into SpeechQuelX/SVG representation respectively as in Figure 4. The SpeechQuelX/SVG files are sent to the corresponding query generator of the target image database. The resulting query for the IBM QBIC image database is shown in Figure 5. The QBIC database will do a search with the abstract sketch as sample and return 5 images as required.

As a closing remark for this section, a SpeechQuel user does not need to construct every abstract sample from scratch. We have designed a group of commands to import a template from a sample library as well as adding a completed sample into the template library. The user can invoke a template as a starting point or even superimpose one template on top of a half-finished sample or another template. This makes the job of orally constructing the abstract sample a lot easier than describing everything from scratch. Also note that the templates are saved as SVG text files, not with the images themselves. It therefore takes up only a small portion of the disk space.

## 5. IMPLEMENTATION AND EVALUATION

In our prototype system, we used the Mandarin-to-text software of the Applied Speech Technologies Inc.(`http://www.speech.com.tw`) as the speech interface. For the image database systems, we employed the QBIC[4] and a locally developed system for their readily accessibility. To evaluate our techniques upon different types of image content, we collected three different image

```
<?xml version="1.0" encoding="BIG5"?>
<ImageDB>
  <Random flag="true" />
  <KeyWord flag="true" keyword="Cats" />
  <Text flag="false" />
  <Color flag="false" weight="0" />
  <Texture flag="false" weight="0" />
  <Shape flag="true" weight="0" />
  <FileNoLimit No="5" />
  <SampleFileName name="sample.jpg" />
</ImageDB>

<?xml version="1.0"?>
<svg width="400" height="200">
<g>
  <rect x="0" y="0" width="401" height="401"
    style="fill:rgb(0,200,0);" transform="rotate(0)" />
  <ellipse cx="214" cy="104" rx="20" ry="13"
    style="fill:rgb(57,57,57);" transform="rotate(90)" />
  <ellipse cx="184" cy="104" rx="20" ry="13"
    style="fill:rgb(57,57,57);" transform="rotate(90)" />
</g>
</svg>
```

*Figure 4.* The SpeechQuelX representation of the query constraints and the SVG representation of the abstract sketch.

```
<?xml version="1.0" encoding="BIG5"?>
<qbic>
  <parameter> -r </parameter>
  <parameter> -T "Cats" </parameter>
  <parameter> -i "sample.jpg" </parameter>
  <parameter> -n 5</parameter>
  <parameter> -f QbDrawFeatureClass </parameter>
</qbic>
```

*Figure 5.* The transformed query constraints for the QBIC database.

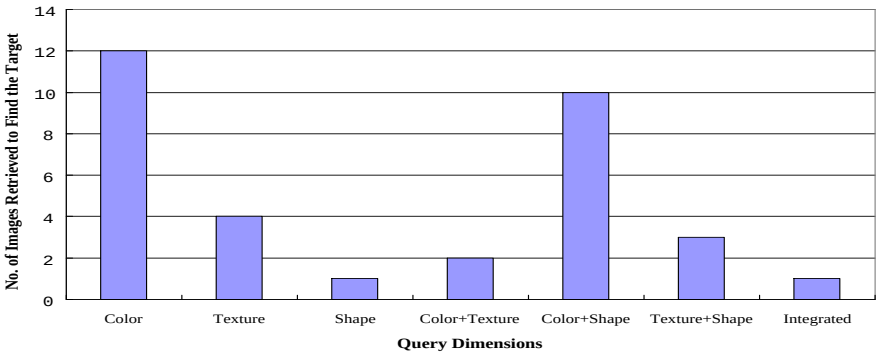


*Figure 6.* Flowers DB Search Effectiveness.

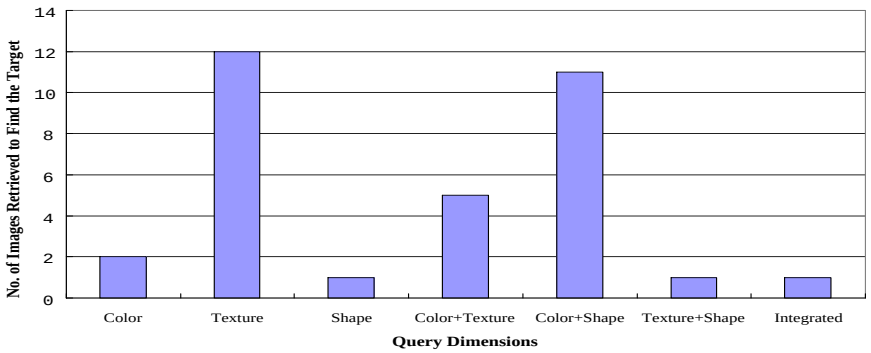


*Figure 7.* Snoopy DB Search Effectiveness.

databases: a flowers gallery, a collection of cat images, and a database of Snoopy cartoon(referred to as **Flowers, Cats,** and **Snoopy** respectively).

## 5.1. Search Effectiveness

In the first set of experiments, we pick an intended target image from each database and use the SpeechQuel language to orally command the system in drawing an abstract sample of the target. We then instruct the system to search the target based on different features such as color, shape, texture, as well as the combination of these query dimensions. The effectiveness of the search is measured by the number of images required for each database to find the target. The results are presented in Figure 6 for the Flowers gallery, Figure 7 for the Snoopy Cartoon collection, and Figure 8 for the Cats database.

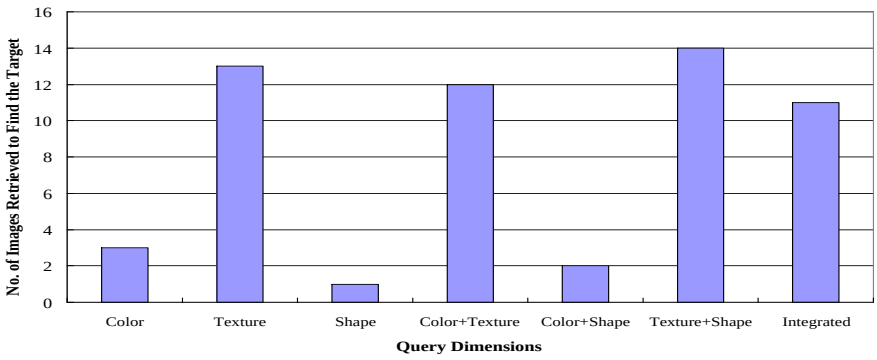


*Figure 8.* Cats DB Search Effectiveness.

It can be observed that different query dimensions have considerably varying effects on three types of image content. For example, query by color is quite effective for both **Snoopy** and **Cats** but fails rather miserably on **Flowers.**This coincides with our expectation since orally discriminating different flowers on color alone is far more difficult than Snoopy and Cats images that are already quite distinctive in their color feature. Among all the dimensions, query by shape consistently outperforms all the others including combination of different features. This is the natural result of our use of simple graphical objects to compose the abstract sketch. It is much more easier to describe the desired image with an abstract sample that mimics the shape of the target than with other features such as texture.

## 5.2. Abstract Sample vs Random Samples

A commonly used retrieval method for many content-based image databases is to provide a set of random samples to start the search. The user then interactively selects an image that is closest to the target as feedback and get another set of similar images in return. The process continues until the intended target image is finally retrieved. It normally takes several rounds of retrieval and feedback to get to the target. One of the most distinctive feature of our approach is to use an abstract image as the starting point. It is therefore our goal in this set of experiments to compare our approach with random sampling.

To make things in favor of random sampling, we use the target image itself as the seed and retrieve a fixed number of similar images(say 3) using various query dimensions. The random sampling process is considered a success as soon as it finds ANY image that falls in the group. It doesn't have to find the exact target image itself. This also reduces the effect of selecting the wrong feedback images which is rather subjective in the first place.

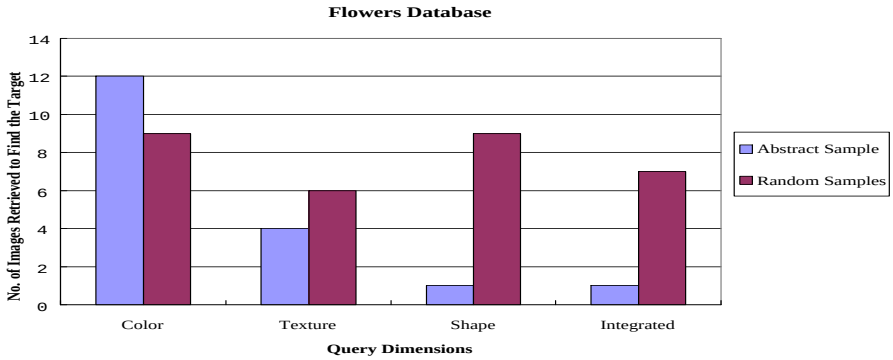


*Figure 9.* Abstract Sample vs Random Samples - Flowers.

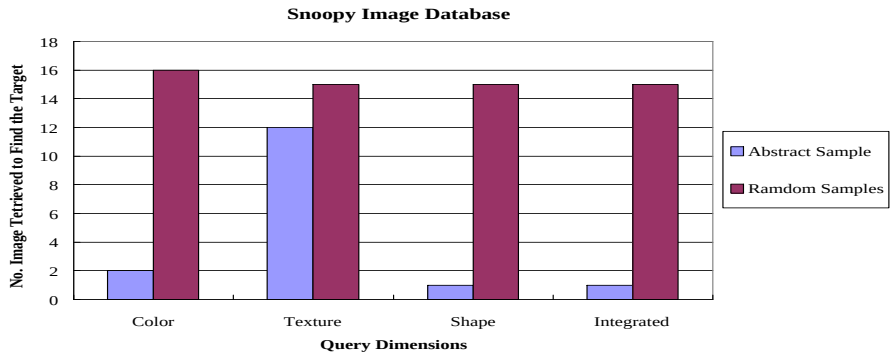


*Figure 10.* Abstract Sample vs Random Samples - Snoopy.

The results of the experiments are presented in Figure 9, Figure 10 and Figure 11. We can see that the abstract sample approach is superior on all three types of images and almost all query dimensions. The only exception is query by color on Flowers which is again not surprising for the difficulty of orally discriminating flowers by color alone. Also note that query by shape is again the most effective and consistent query dimension.

## 5.3. Expert vs Naive Users

As stated earlier, the effectiveness of the traditional interactive retrieval method dependents heavily on the user of picking the most appropriate query dimension(s) and making the right feedback choice(s). This is not an easy job for a naive user who may not have any knowledge about content-based image retrieval. Our approach is much more user friendly in that even a naive user

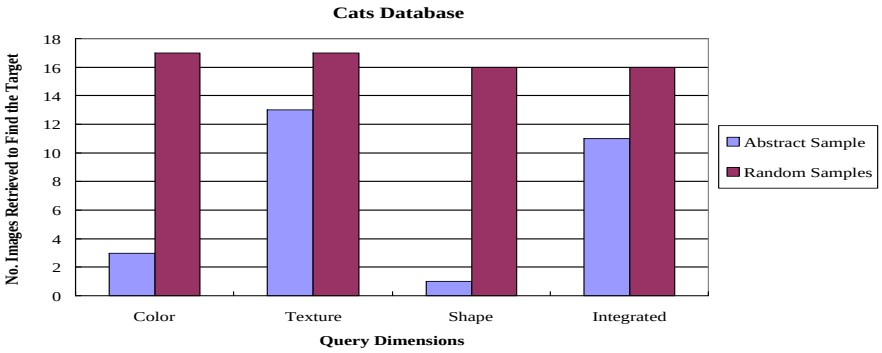


*Figure 11.* Abstract Sample vs Random Samples - Cats.

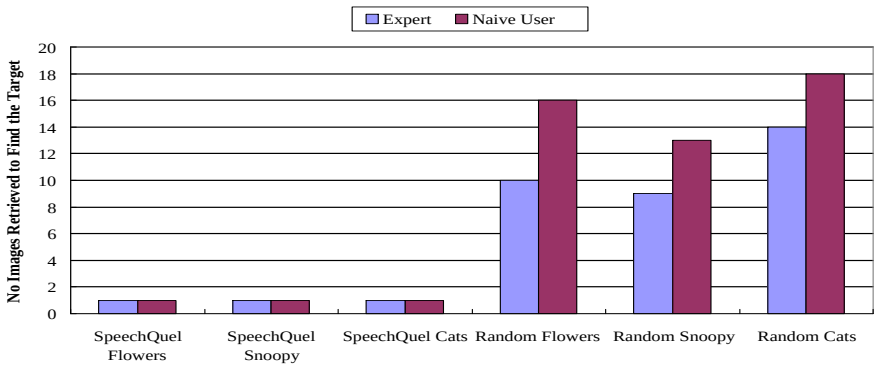


*Figure 12.* Expert vs Naive Users.

can describe an abstract sample that is equally effective as an expert. For the purpose of verifying our claim, we have conducted a set of experiments to compare the performance of expert vs. naive users using both SpeechQuel and the traditional interactive approach. The result (Figure 12) is very encouraging. While experts consistently outperform naive users using traditional method, both are equally effective using our approach. This is a clear indication that Oral-Query-by-Sketch is a quite natural and easy to learn method that successfully raises the performance of naive users to a higher level that is comparable to that of the experts.

## 5.4. Search Efficiency and Translation Overhead

In this set of experiments, we intended to measure the system efficiency in terms of search time as well as the percentage of time used for data transfor-

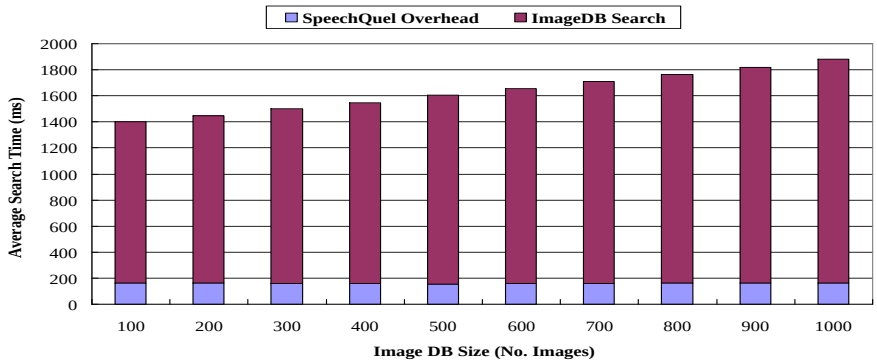


*Figure 13.* Search efficiency and translation overhead - query by shape.

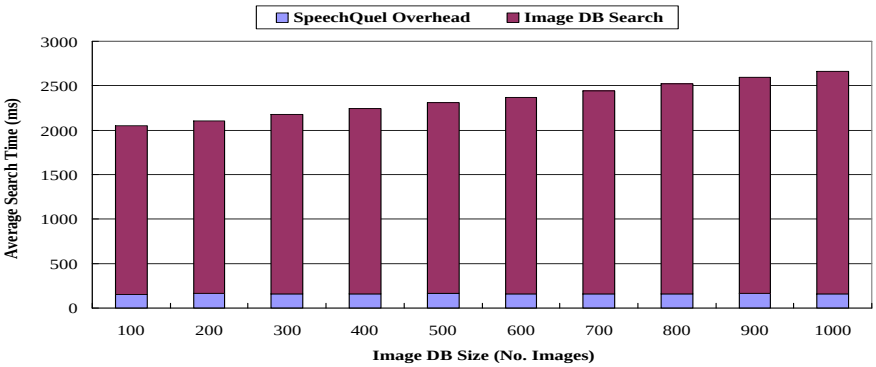


*Figure 14.* Search efficiency and translation overhead - query by color, shape and texture.

mation. To be an effective technique, the translation overhead must be kept to a minima. We measured the system response time with database size ranging from 100 to 1000 images using both query by shape and by the combination of color, shape and texture. The result is again quite encouraging as presented in Figure 13 and Figure 14. We can see that, the total search time increases linearly with database size which is not surprising. What's more interesting is that the translation overhead takes up a small and fixed amount of cost that is independent of the size of the database and the query dimension of choice. This is highly desirable for a speech access method both in normal usage and in demanding situations such as mobile environments.

As a summary, the Oral-Query-by-Sketch framework enables speech access to image databases with high precision and low overhead. The benefits of hand-free access, low image transmission requirement and low overhead also make it highly suitable for disable users or for mobile multimedia access.

## 6. CONCLUSIONS AND FUTURE WORK

We have proposed and implemented the Oral-Query-by-Sketch framework for integrated access to multiple image databases. The use of abstract sketch and speech access make it especially useful for situations that demand hand-free operations. Empirical results successfully demonstrate not only the feasibility but also the effectiveness of our approach. We plan to extend and migrate our prototype system onto a PDA and wireless communication platform. There is also a need to construct a rich template library to help the user in constructing the abstract sample. We are also evaluating the possibility of automatically constructing a set of templates from a given image source. Since the abstract sketch described by the user is typically quite different from the target image, there is a potential benefit of developing effective matching criteria and content-based image retrieval methods to explore the similarity on an abstract level.

## References


[1] Y. Aslandogan and C. Yu. Techniques and systems for image and video retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):56–63, January/February 1999.

[2] S.-K. Chang and A. Hsu. Image information systems: Where do we go from here? *IEEE Trans. on Knowledge and Data Eng.*, 4(5):431–442, Oct. 1992.

[3] D. Connolly. Extensible markup language (xml). w3c (mit, inria, keio), 1999. Available at: http://www.w3.org/XML/.

[4] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, and Q. H. et al. Query by image and video content: The qbic system. *IEEE Computers*, 28(9):23–32, Sept. 1995.

[5] T. S. Huang, S. Mehrotra, and K. Ramchandran. Multimedia analysis and retrieval system (MARS) project. Proc. 33rd Annual Clinic on Library Application of Data Processing - Digital Image Access and Tetrieval, 1996.

[6] C. Lilley. Scalable vector graphics (svg). w3c (mit, inria, keio), 2000. Available at: http://www.w3.org/Images/SVG/Overview.html.

[7] W. Y. Ma and B. S. Manjunath. A comparison of wavelet transform features for texture image annotation. In *Proc. IEEE Int. Conf. On Image Proc.*, 1995.

[8] S. Muller, S. Eickeler, and G. Rigoll. Multimedia database retrieval using hand-drawn sketches. In *ICDAR '99. Proceedings of the Fifth International Conference on Document Analysis and Recognition*, pages 289–292, 1999.

[9] Y. Rui, T. Huang, and S. Chang. Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10(4):39–62, Apr. 1999.

[10] E. D. Sciascio and M. Mongiello. Query by sketch and relevance feedback for content-based image retrieval over the web. *Journal of Visual Languages and Computing*, 10(6):565–584, 1999.

[11] F. Smith, B. Peters, D. Crookes, and G. Philip. Speech access to a document database system. In *Proc COMAD'89, Hyderabad, India*, Nov. 1989.

[12] J. Wang, W.-J. Yang, and R. Acharya. Color clustering techniques for color-content-based image retrieval from image databases. In *Proc. IEEE Conf. on Multimedia Comput. and Syss*, 1997.

[13] Y. Zhong and A. K. Jain. Object localization using color, texture and shape. *Pattern Recognition*, 33(4):671–684, Apr 2000.