Multimodal Image Search

[1]Sushmita N. Bankar, [2]Prof. D. B. Kshirsagar
[1,2]Computer Engineering Department,
College of engineering, Kopargaon, India

***Abstract* – The system describes a novel multimodal interactive image search on mobile devices. The framework, the Joint search with ImaGe, Speech, And Word Plus (JIGSAW+)[7], exploits the multimodal information and characteristic client co-operations of cell phones. It is intended for clients who as of now have pictures in their minds yet have no exact depictions or names to address them. By depicting it using speech and after that refining the perceived question by intuitively creating a visual inquiry using exemplary images, the client can without much of a stretch discover the desired pictures through a couple of regular multimodal connections with his/her cell phone. The performance of the system is tested in terms of precision and recall, and it is observed that better precision and recall is obtained. Also the CBIR results for Image-to-image search module are better.**

***IndexTerms* – Mobile visual search, multimodal search, interactive search, mobile device.**

I. INTRODUCTION

Now-a-days image search is an intriguing issue in both machine vision and information retrieval with numerous applications. The conventional desktop image search frameworks with text queries have ruled the client conduct for a long stretch. Contrasted and content pursuit, guide hunt, and photograph to-inquiry, visual (picture and feature) inquiry is still not extremely famous on the telephone, however image search has turned into a typical instrument on the PC since 10 years prior, with which the client can include content question to recover pertinent pictures. A primary motivation behind why such picture seek applications are not prominent on cell phone is that the current picture search applications don't splendidly suit to the portable and nearby situated client aim. Because of this, the indexed lists are seldom valuable and the client encounter on the telephone is not generally pleasant. Most importantly, writing is a dull employment on the telephone regardless of whether a little console or a touch screen is utilized. Despite the fact that voice queries are accessible on a few gadgets, there are still numerous cases that semantic and visual aim can barely be communicated by these descriptions for search. For instance, in a typical picture pursuit undertaking, the client may have officially imagined the general thought of expected pictures, for example, color designs and arrangements. Nonetheless, the client normally needs to get perfect pictures in the midst of substantially more unimportant results.
A small screen limits the presentation of searching results, which requires the top results to be more relevant while on the phone. However, using only text as search query can hardly meet this end. The surrounding texts of the web images are not always correct. Even the tags of the some human-labeled datasets such as Flickr images are unreliable. Moreover, on the one hand, the user must know the exact terms the annotator used in order to be able to retrieve the images he wants. On the other hand, textual annotations are also language-dependent. Actually, there are more images which have no text information on the web repository. All this deficiency can ruin a good user experience of text-based image search system on the mobile phone.

II. LITERATURE SURVEY

Bernd Girod et al. gave an extensive outline of photograph to-search in [1], from the structural planning of an effective versatile framework to the system of a image recognition algorithm. Robust local image features achieve a high degree of invariance against scale changes, rotation, as well as changes in illumination and other photometric conditions. The BoW approach offers resiliency to partial occlusions and background clutter, and allows design of efficient indexing schemes.
In case of visual representation of images, the images are over segmented before color feature extraction using algorithm by Felzenszwald and Huttenlocher [2]. In this paper, authors define a predicate for measuring the evidence for a boundary between two regions using a graph-based representation of the image. They then develop an efficient segmentation algorithm based on this predicate, and show that although this algorithm makes greedy decisions it produces segmentations that satisfy global properties.
On Flickr [3], everyone gets 1000GB of free storage, enough space for more than 500,000 photos. Their powerful search technology means you can find them anytime you want. Its the world's largest photography community.
Affinity Propagation (AP) algorithm [4] is received to group the applicant pictures into a few gatherings. Frey and Dueck devised a method called affinity propagation, which takes as input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges.
Speech recognition is now a much more mature technique than image recognition. Especially for speech to text technique, the state-of-art accuracy achieves an accuracy of 98% in quiet environment using Hidden Markov Models [5]. Hidden Markov Models (HMMs) provide a simple and effective framework for modeling time-varying spectral vector sequences. As a consequence, almost all present day large vocabulary continuous speech recognition (LVCSR) systems are based on HMMs.
Tuytelaars, Bay and Van Gool presents a novel scale and rotation invariant detector and descriptor, coined SURF (Speeded-Up Robust Features) [6]. SURF approximates or even outperforms previously proposed schemes with respect to repeatability,

distinctiveness, and robustness, yet can be computed and compared much faster. This is achieved by relying on integral images for image convolutions; by building on the strengths of the leading existing detectors and descriptors; and by simplifying these methods to the essential. This leads to a combination of novel detection, description, and matching steps.
In paper [7], a multimodal mobile search framework is designed intended to do visual search. In [10] the creators fabricate a Sketch2photo framework that uses basic content explained line representation to naturally combine practical pictures. They additionally utilize content and representation to hunt down formats which are then sewed on a foundation to create a montage. Then again, their work concentrates on picture making rather out of picture retrieval.
As the speech recognition got to be develop, telephone applications utilizing speech recognition quickly becomes as of late. The most illustrative application is Apple Siri [9], which joins speech recognition; natural language understanding and learning based searching procedures.
In scholastic circles, there is not an enormous contrast from industry. Researchers for portable pursuit likewise concentrate fundamentally on photograph to-inquiry methods. Conventional features, for example, SIFT [8] and Speeded Up Robust Feature (SURF) [6] are broadly utilized as a part of such visual search frameworks due to their invariance to illumination, scale and rotation. Chandrasekhar et al. talked about their compression and in addition proposed another feature of Compressed Histogram of Gradients (CHoG) [12]. It can quantize and encode inclination histogram with Huffman and Gagie trees to deliver low bit-rate descriptors. Furthermore, different frameworks with improved system and better indexing are produced to scan for milestones, books, CD covers [11] and so forth. In [10], diverse local descriptors are analyzed in an arrangement of CD cover inquiry. SIFT is generally acknowledged as the best performed peculiarity and CHoG has preference in low-bit transmission. Different methods are additionally utilized as a part of visual inquiry, for example, barcodes and OCR.
In both industry and academic circles, it is found that there are few works on mobile image search. As a result, the JIGSAW+ system differs to existing mobile visual search systems in that it represents multiple search modes in one system by which the mobile users also can provide images as a query to search for resulting images on the go.

III. SYSTEM ARCHITECTURE

With rich interactions and visual techniques, a natural image search system, Joint search with ImaGe, Speech, And Word Plus (JIGSAW+) is introduced to deal with the aforementioned scenarios which enables the user to conduct a image search with visual aids. The objective is to design a efficient visual aided image search application on mobile phone combined with local spot and scene search. Fig.1. illuminates the architecture of the multimodal search system.

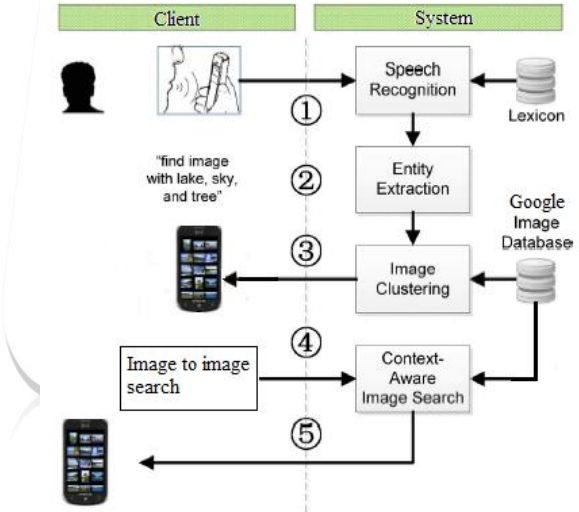

Fig.1 Overall architecture of JIGSAW+

The search procedure of this system consists of the following phases:
1. The user enters a text query or speaks a natural sentence to describe the intended images,
2. The speech is recognized and further decomposed into keyword(s) which can be represented by a text query,
3. Clustering algorithm is applied on the images retrieved from google to get more relevant results,
4. The user can also give image as a query, and
5. Image search is done according to the contents of the image and results are displayed to the user.

Implementation will be done according to the following major modules:

1. Text-to-search Module:
In this module we have to give a text as a description of an image so that a search is performed on that description and result is displayed on the screen. We have to type a text on the given text box and click on search button and the result is displayed on grid view provided by application. Also we can view the full screen images.

2. Voice-to-search Module:

The whole search interaction begins with the natural language understanding. In this system, a simple entity extraction strategy will be adopted to handle this problem. The entity extraction from speech can be divided into two steps: 1) speech recognition, and 2) key entity extraction.
A Hidden Markov Model (HMM) has the capacity to handle both natural sentences and phrase fragments. Typically, such word extraction is carried out by either key word extraction or unimportant word filter.
Clustering algorithm is applied on the images retrieved from the google in case of text and voice modules to obtain more relevant results. For this task, Affinity Propagation (AP) algorithm is used.
The Affinity Propagation algorithm takes as input a collection of real-valued similarities between data points (matrix S), denoted by s(i; j), which indicates how close (similar) the two nodes are. Here, similarity is calculated using eq. 7. When i = j; s(i; j) will store the probability that this i point will be selected as ”exemplar”. The point k with higher s(k; k) value is more likely to be chosen as ”exemplar”.
Real-valued messages will be exchanged between data points until a good set of exemplars and patterns emerge, based on the inputs. There are two kinds of messages, “responsibility” r(i; j) and “availability” a(i; j). The “responsibility”, r(i; j), which will be sent from data point i to candidate exemplar point j (j is still a data point there), reflects how suitable it is that point j serves as the exemplar for point i, taking into account other potential exemplar for i. The “availability”, a(i; j), sent from candidate j to point i, which will reflect that how proper it is that point i choose point j as its exemplar, considering support from other point to choose j as exemplar. In each iteration, the algorithm will update these messages until the convergence conditions get satisfied. In other words, we will have another two matrices A and R, and we will update the matrices in every iteration.

Algorithm: Affinity Propagation (AP)
1. At first, all responsibilities will be set to zero, a(i, k) = 0, for any i, k. Then, begin iteration, update matrix R, the formula is r(i, k) = s(i, k) - max(a(i, k’) + s(i, k’)) where k’ != k.
2. The next step is updating matrix A, there are two cases:
when i != k; a(i, k) = min(0; r(k, k)+f(k)),
where f(k) = $\sum$ max(0; r(i’, k)), and i’ $\notin$ {i, k}
when i = k; a(k, k) = f(k); f(k) is the same as above.
3. From the matrix R and A, for every point i, find the point k that maximizes a(i, k) + r(i, k). Here k will be the exemplar for point i, and if i = k, point i itself will be a exemplar.
4. Do the iterations until some conditions satisfied.

3. Image-to-search Module:
In this module we have to give an example image to search similar type of images.
To give example image either we have to select image from SD card or from application database where selected images are stored by user.
Here input image will be processed and some low level features will be extracted from the image. Like color feature extraction and texture feature extraction. Then feature vectors will be formed and then based on these feature vectors, similarity measures will be calculated.
With the vector representation, the cosine similarity is calculated between each pair of images:

$$sim(f1, f2) = \frac{(f1, f2)}{\sqrt{(f1, f1)(f2, f2)}} \quad \ldots(1)$$

1. Color Feature Extraction:
Color is the most extensively used visual content for image retrieval. Its three-dimensional values make its discrimination potentiality superior to the single dimensional gray values of images [14].
Color Coherence Vector:
JIGSAW+: Joint search with ImaGe, Speech, And Word Plus. It is a histogram-based method for comparing images that incorporates spatial information. It classifies each pixel in given color bucket as either coherent or incoherent, based on whether or not it is part of large similarly-colored region.

2. Texture Feature Extraction:
Texture is another important property of images. Various texture representations have been investigated in pattern recognition and computer vision [15]. Basically, texture representation methods can be classified into two categories: structural and statistical. Structural methods, including morphological operator and adjacency graph, describe texture by identifying structural primitives and their placement rules. They tend to be most effective when applied to textures that are very regular.

Texture Element Feature Characterization:
It is a structural method for texture analysis of CBIR. It uses local pattern as key and is expected to return more relevant images in CBIR when used as feature. Here, Local Patterns are characterized by using two methods:

Mathematical Model
The computational complexity of this implementation is a P class.
Let the main input set (It) contains three variations of inputs. Ex. Text (T), voice (V) and image (I).
I_t = {T, V, I}

There is an input set of text input as T where t1, t2, …, tn are extracted keywords.

T = {t1, t2,…, tn}

There is an input set of voice as an input, represented as V. Where w1, w2,…, wn are recognized words through voice recognition.

V = {w1, w2,…, wn}

Now there is an input set of image as an input, represented at I. Where i_1, i_2,…, i_n are different main object names from that image.

I = {i_1, i_2,…, i_n}

Together with text Tq and exemplary image Iq, the visual exemplar is a triplet represented as,

Cq = {Tq, Iq, Rq}

$$Q = \{C_q\}_q^k = 1$$

The whole query (Q) is a set of k exemplars.

Processes:

Consider a set of important processes which are used in this system.

P1 = Speech Recognition

P2 = Entity Extraction

P3 = Image Search

P4 = Gallery image selection

Output:

O1 = Grid of Images

Process State Diagram:

Here, initial process, process P1, P2, P3 and process P4 are denoted by q0, q1, q2, q3 and q4 respectively. Refer figure 2.

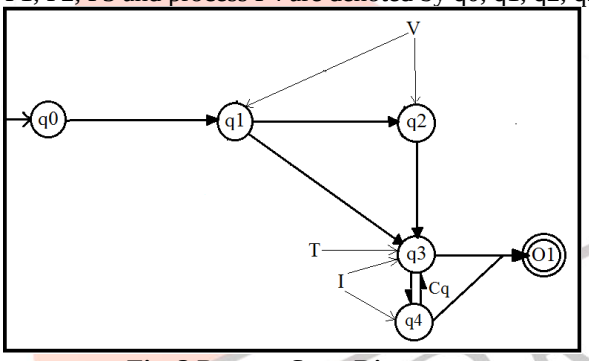

Fig. 2 Process State Diagram

IV. RESULTS AND DISCUSSION

The system with phone application is built to test the multimodal visual search system. The back-end of the test system is deployed on a remote server, and a front-end interface application is developed on an Android phone with android version 4.2 and above.

Following fig. 3 shows the interface of the app on the phone. The application is called JIGSAW+ which means a multimodal visual search system "Joint ImaGe, Speech, And Word Plus." Figure 4 and 5 shows the results for Text-to-image search and Voice-to-image search module.

Compared to text input, natural sentences are given as an input in case of voice query. Due to such voice queries, search results are improved significantly. This can be observed in following result tables (Table 1, 2 and 3). Here, precision is calculated from the following formula:

$$precision\,(\%) = \frac{No.of\,Relevant\,Images}{Total\,No.of\,Images\,Retrieved} \times 100 \quad \ldots(2)$$

In case of next module i.e. Image to search, precision will be improved as compared to these two modules.

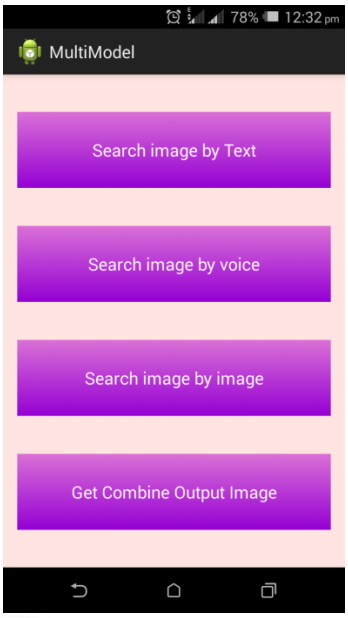

Fig.3 Start page of the multimodal search app

Fig.4 Text to Image search results screenshot

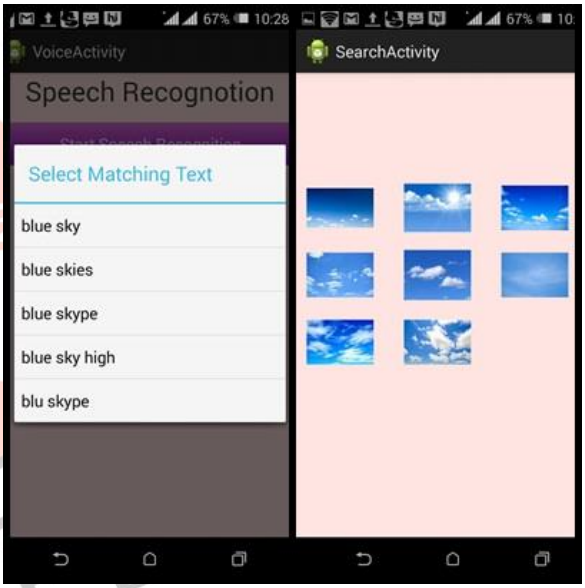

Fig. 5 Voice to image search module results

Table 1: Results obtained from text query set 1

sr. no	query	no. of query entities	no. of relevant o/ps	precision out of 8 (%)
1	glass	1	8	100
2	glass juice	2	8	100
3	glasses juice	2	7	87.5
4	glass juice red	3	7	87.5
5	glass red juice	3	8	100
6	glass mango juice table	4	5	62.5

Table 2: Results obtained from text query set 2

sr. no.	query	no. of query entities	no. of relevant o/ps	precision out of 8 (%)
1	home	1	8	100
2	home trees	2	8	100
3	home grass	2	8	100
4	red home grass	3	3	37.5
5	home grass mountain	3	6	75

Table 3: Results obtained from voice query set 1

sr. no.	query	Voice recognition	no. of query entities	no. of relevant o/ps	precision out of 8 (%)
1	glass	done	1	8	100
2	home	done	1	8	100
3	juice glasses	done	2	7	87.5
4	glass with mango juice	done	2	8	100
5	red house with grass	done	2	8	100

From the above text-to-image result tables 1 and 2 we can observe that, to search images including more objects, user has to provide a query with that much entities. So the complexity of input query goes on increasing. Also the precision here is calculated for upper 8 images. As the value of upper 'k' images goes on increasing, precision value goes on decreasing.
Instead in case of voice input, user can speak natural sentence to give a search query which is not that much difficult task. So the typing task of the user is reduced. For precision related to voice-to-image search, Refer table 3. In case of voice-to-image search, precision remains same as that of the text-to-image search.
Finally for image to image search, user just have to provide image from gallery as an input. And the relevant images are displayed to the user. Here the user experience is also better in case of input query and the precision related to the search results is also better. Refer table 4 and 5.

Table 4. Precision for upper 'k' retrieved images for 'horse' image as an input

Upper 'k' retrieved images	Precision/100 retrieved images
10	100
20	100
30	86.66
40	85
50	80
60	80
70	77.14
80	70
90	63.33
100	59

Table 5. Precision for upper 'k' retrieved images for 'Bus' image as an input

Upper 'k' retrieved images	Precision/100 retrieved images
10	100
20	100
30	100
40	97.5
50	96
60	90
70	85.71
80	81.25
90	78.88
100	77

V. CONCLUSION

Subjective experiment shows that JIGSAW+ is an effective complementary tool to existing mobile search applications, especially in cases where users have only partial visual clues in their minds. Compared to text-based retrieval system the performance of the JIGSAW+ is boosted. Search relevance in improved in visual input mode than the text to image search. The user's search experience on mobile device is thus significantly improved by this game-like image search system. More relevant results are obtained in case of visual search i.e. Image-to-image search module.

VI. ACKNOWLEDGMENT

I take this opportunity to express my profound gratitude and indebtness to my guide and H.O.D. Prof. D. B. Kshirsagar, Department of Computer Engineering, S.R.E.S. C.O.E. Kopargaon; for his personal involvement and constructive criticism provided beyond technical guidance during the project work and the course of the preparation of the report. He has been keen enough for providing me with the invaluable suggestions from time to time. Above all, his keen interest in the project helped me to come out with the best. I would also like to thanks Prof. P. N. Kalavadekar, ME Coordinator, Collage of Engineering, Kopargaon; for providing necessary lab facilities during the period of seminar work.
I would like to thank my parents and my friends who have constantly bolstered my confidence and without whose moral support and encouragement, this seminar would have been impossible. This paper is the outcome of the help of these. Any error that might have crept in is solely mine.

REFERENCES

[1]. D. Chen, N. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. Tsai, B. Girod, V. Chandrasekhar and R. Vedantham, "Mobile visual search", IEEE Signal Process. Mag., 2011.
[2]. P. Felzenszwalb, and D. Huttenlocher, "Efficient graph-based image segmentation", Massachusetts Institute of Technology, Cornell University, 2004.
[3]. Flickr. http://www.flickr.net/.
[4]. B. Frey and D. Dueck, "Clustering by passing messages between data points", Department of Electrical and Computer Engineering, University of Toronto, 10 Kings College Road, Toronto, Ontario M5S 3G4, Canada., 2007.

[5]. M. Gales and S. Young, ”The application of hidden markov models in speech recognition”, Foundations and Trends in Signal Processing, Vol. 1, 2008.
[6]. T. Tuytelaars, H. Bay and L. Van Gool, ”SURF: Speeded-up robust features”, Proc. ECCV, Belgium, 2008.
[7]. Tao Mei, Jingdong Wang, Houqiang Li, Yang Wang and Shipeng Li, ”Interactive multimodal visual search on mobile device”, IEEE transactions on multimedia, 2013.
[8]. David G. Lowe, ”Distinctive image features from scale-invariant keypoints”, International Journal of Computer Vision, 2004.
[9]. Siri. http://www.apple.com/iphone/features/siri.html.
[10]. P. Tan, A. Shamir, T. Chen, M.-M. Cheng and S. M. Hu, ”Sketch2photo: Internet image montage”, 2009.
[11]. A. Lin, G. Takacs, S. S. Tsai, N. M. Cheung, Y. Reznik, R.Grzeszczuk, V. Chandrasekhar, D. M. Chen and B.Girod, ”Comparison of local feature descriptors for mobile visual search”, Proc. IEEE Int. Conf. Image Process., 2010.
[12]. D. Chen, S. Tsai, R. Grzeszczuk, V. Chandrasekhar, G. Takacs and B. Girod, ”CHoG: Compressed histogram of gradients a low bit-rate feature descriptor”, in Proc. IEEE Conf. Comput. Vis. Pattern Recogn, 2009.
[13]. J. Wang, H. Li, Y.Wang, T. Mei and S. Li, ”JIGSAW: Interactive mobile visual search with multimodal queries”, in Proc. ACM Multimedia, 2011.
[14]. Greg Pass, Ramin Zabih, Justin Miller, ”Comparing Images Using Color Coherence Vectors”, Computer Science Department, Cornell University, Ithaca, gregpass,rdz,jmiller@cs.cornell.edu
[15]. K. Jalaja, Chakravarthy Bhagvati, B. L. Deekshatulu, Arun K. Pujari, ”Texture Element Feature Characterizations for CBIR”, Dept. of Computer and Information Sciences, University of Hyderabad, Hyderabad.