USABILITY EVALUATION FOR IMAGE RETRIEVAL BEYOND DESKTOP APPLICATIONS

Thomas Käster[1], *Michael Pfeiffer*[2] *and Christian Bauckhage*[3]

[1]Vis-à-pix GmbH, Potsdam, Germany
[2]Bielefeld University, Bielefeld, Germany
[3]Deutsche Telekom AG, Laboratories, Berlin, Germany

ABSTRACT

Interactivity is a key concept in modern content-based retrieval. Therefore, in addition to the ability to learn from user generated data, easy and intuitive to use interfaces are an important area of research in (multi)media retrieval. In this contribution, we focus on the latter aspect and present how different modalities like speech and gestures on super sized touch screen facilities may be integrated to accomplish the goal of intuitive interaction. In order to evaluate our approach, we conducted a series of usability experiments. Their results demonstrate that our multimodal user interface allows for both, comfortable and successful interactive image retrieval.

1. INTRODUCTION AND MOTIVATION

In the last decade, smart rooms have become a popular topic of intelligent systems research. MIT's Intelligent Room [1] and Microsoft's EasyLiving [2] are just two examples of projects targeting the home environment of the future. They all augment the traditional home environment with technologies which originated in different fields of computer science. Computer vision, speech processing and related approaches to human-machine interaction are applied to create an environment that provides innovative and comfortable living conditions. As a major design principle of high-tech housing spaces [1], living with advanced technologies should not be intrusive. The possibility for easy, intuitive and seamless interaction with the environment is a prerequisite of intelligent rooms.

Dealing with a different aspect of future living, a recent user study by Eggen et al. [3] revealed that photographs are among the most important objects of a living space. They found that retrieving photos and sharing visual memories with others plays a key role in current humans social life. To this end and due to the ever growing amount of digital images, efficient navigation in photo libraries is an essential element of tomorrow's living.

In this paper, we consider this aspect of ubiquitous image retrieval and image database access from the point of view of intelligent interfaces and smart room technologies. We briefly sketch the key components and architecture of a content-based image retrieval (CBIR) system we developed in earlier work. Afterwards, we describe our approach to multimodal interaction. As shown in Fig. 1, using speech and gestures on a wall-mounted touch screen allows for easy and intuitive image browsing. Then we present an extensive usability study on multimodal, interactive image retrieval. Finally, a conclusion will end this contribution.

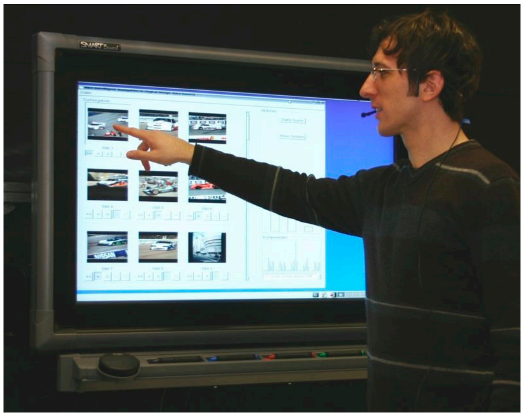


Fig. 1. Interactive CBIR using speech and gestures.

2. SYNOPSIS OF THE INDI SYSTEM

Our content-based image retrieval system INDI results from a project on techniques for **I**ntelligent **N**avigation in **D**igital **I**mage Databases [4, 5]. Figure 2 shows the architecture of the system; its main features are low-level image representations, adaptivity, and facilities for multimodal interaction. Following an idea of Rui and Huang [6], retrieval is done in a hierarchical manner. It applies low-level features such as color histograms or texture descriptors. However, we do not only consider global image features but also extract local descriptors for salient image regions [5]. Since the system relies on the common query-by-example paradigm, the user initially selects an image that fits his search intention from a random set of samples. The query image is then compared to

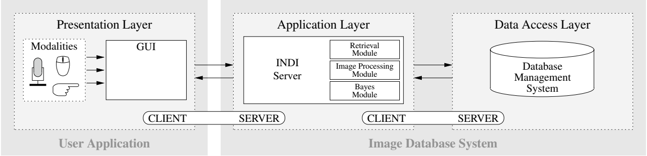


Fig. 2. The INDI system consists of a multimodal GUI, the central retrieval server and a database management system.

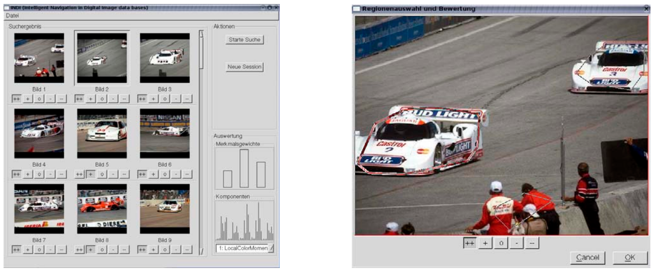


Fig. 3. Screen shot of the graphical user interface with the main window, left, and the region selection window, right.

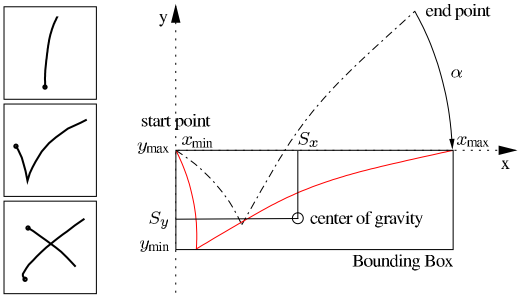


Fig. 4. Touch screen gestures and feature extraction.

all images in the database and the most similar ones are displayed to the user. To this end, each image and the query image are compared in the different feature spaces. The resulting dissimilarity values are combined into an overall value. Even though hierarchical CBIR facilitates handling of huge data sets, it suffers from the semantic gap between the user's high-level interpretation of an image and the low-level image descriptors. State of the art systems therefore rely on the concept of the *human-in-the-loop* to bridge this gap [6, 7]. Correspondingly, our system asks the user to rate the results of each retrieval step which allows for continuous parameter tuning and adaption to the user's search intention.

3. MULTIMODAL INTERACTION

This sections describes how to use natural modalities for easy and intuitive interaction in image retrieval with devices other than desktop computers.

3.1. Mouse, Touch Screen and Touch Screen Gestures

The most common way of interacting with the graphical user interface (GUI) of our system is to use a mouse and to rely on the conventional metaphor of buttons and sliders (see Fig. 3). In addition, our system can be operated from a touch screen that supports mouse-like actions. Unfortunately, emulating the right mouse button is impossible with conventional touch screen hardware and double-clicks may suffer from imprecision. We therefore introduce several touch screen gestures to provide more comfortable means for human-system interaction. Currently, our system supports three different gestures (see Fig. 4). They allow for enlarging images (*stroke*) or selecting examples for retrieval (*hook* or *cross*).

Gesture recognition is done by means of a polynomial classifier. Since this requires a vectorial representation, gesture trajectories are transformed into five dimensional representations; Fig. 4 exemplifies this. First, the trajectory is translated so that its starting point coincides with the origin of the coordinate system. Then, it is rotated by an angle α so that its endpoint comes to lie on the x-axis. The value $v_P = y_{\text{max}}/(y_{\text{max}} - y_{\text{min}})$ characterizes the gesture's height above the x-axis where y_{min} and y_{max} denote the minimal and maximal y-coordinates of its bounding box. Similarly, the length of the trajectory is scaled by the length of the diagonal of the bounding box leading to a value v_L, Together with the center of gravity of the trajectory, these values form a vector

$$\vec{r}_{\text{gesture}} = (\alpha, v_P, v_L, S_x, S_y)^T.$$

3.2. Speech Processing and Linguistic Referencing

Our system may also be operated using natural speech. To this end, we adopt a speech recognition module from a toolbox developed by Fink [8]. Its main component is a statistical speech recognizer based on Hidden-Markov-Models. Together with a parsing component that exploits grammatical restrictions already in the recognition phase, it yields a better accuracy than most conventional speech recognition systems [9]. Integrated into our system, it allows for using verbal commands to trigger actions such as selecting the initial example image or rating the images of a result set.

Our system supports searching for similar images as well as searching for similar image regions. For seamless interaction, it is therefore desirable that parts of an image can be referenced using natural phrases which may contain prepositions, adjectives and comparisons. For instance, the car on the left in the region selection window in Fig. 3 may be rated by saying something like *"the bright region on the left is very good"*. Mapping this utterance onto appropriate GUI action requires linking the speech signal to visual characteristics of detected image regions.

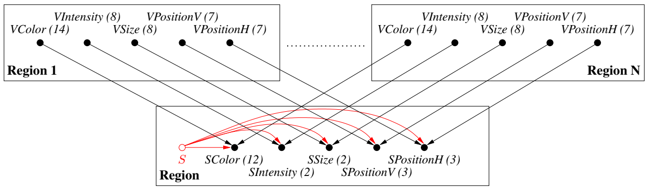


Fig. 5. A Bayesian network for image region description.

In order to overcome the common but invalid assumption of infallible speech recognition and one-to-one mappings between the results of speech and image processing, we follow an approach by Wachsmuth and Sagerer [10]. Here, fusing speech and vision is treated as a probabilistic decoding process which is modeled using Bayesian networks. This is preferable since prepositions like *"left of"* or adjectives like *"bright"* as in the above example are inherently fuzzy. In the Bayesian framework, we represent each region described in an utterance and each region detected in an image by means of separate subnetworks. As illustrated in Fig. 5, we consider the region attributes 'color', 'intensity', 'size' and 'position' (horizontal and vertical). The numbers in the parentheses indicate the dimensions of the corresponding random variables. Given a set of conditional probability tables and the currently observed evidences e of the subnetworks, the verbally referenced image region is determined from probabilistic inference. A selection variable $S = \{1, , 2, \ldots, N\}$ is central for computing the necessary conditional probabilities. For instance, the conditional probability of node $SColor$ is defined by:

$$P(SColor|VColor_1, \ldots, VColor_N, S) = \begin{cases} P(SColor|VColor_1), & \text{if } S = 1 \\ \vdots & \\ P(SColor|VColor_N), & \text{if } S = N \end{cases}$$

The maximum a-posteriori hypotheses of S defines the most probable image region referenced by the current utterance:

$$r^* = \underset{r \in \{1,2,\ldots,N\}}{\operatorname{argmax}} P(S = r|e)$$

3.3. Combining Speech and Gesture

In addition to an isolated usage of a modality, in natural interaction, there are many occasions where several modalities are used simultaneously. For instance, while uttering commands such as *"enlarge this image"*, many users underline their intention by pointing to the image. In order to initiate the appropriate GUI action, our system therefore has to fuse asynchronous events like this. Consequently, a special event handler process all input events. Events that can be mapped directly to a suitable GUI action are forwarded to the main GUI handler. All other events are forwarded to event fusion queues; thus speech and gesture inputs are combined only if they occur within a certain time interval.

4. EVALUATING THE INTERACTION

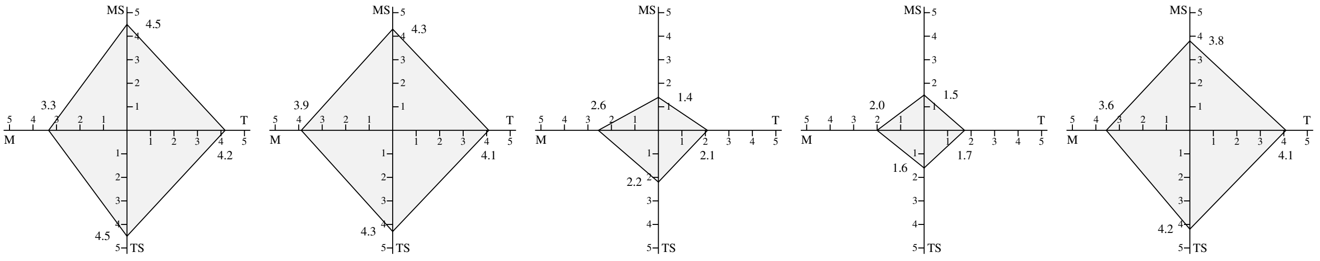
In this section, we present a comprehensive user study of our multimodal retrieval interface. It was based on experiences from evaluating an earlier version of the system [5]. This time, more subjects took part and more variables contained in the recorded interaction data were analyzed and charted.

We considered a database of 1250 images from the ArtExplosion collection showing scenes from 10 semantic classes such as "sunsets" or "car racing". A total of 40 individuals (7 female and 33 male) with different academic education, e.g. teaching, psychology, law or computer science were asked to interact with our system. They were aged between 19 and 37 and did not have any experience in content-based image retrieval. To determine how the input modality affects retrieval results or user satisfaction, the subjects were divided into four groups. The first group was restricted to the *mouse* (M) while interacting with the system. The second group interacted by using the *touch screen* (T); the remaining groups relied on *mouse and speech* (MS) and *touch screen and speech* (TS), respectively. Each group had to accomplish three target searches. The goal was to retrieve a specific image from the database within a time limit of three minutes. If it was exceeded, the experiment was counted as a failure.

Diverse technical data were recorded for a quantification of interaction quality. These include the average time the subjects needed for a task as well as the average number of GUI actions (e.g. moving a slider or pushing a button) they performed. In addition, our subjects were handed a questionnaire which was devised with help from colleagues in psychology. As in our earlier study [5], it focused on criteria adopted from Preece et al. [11]: the *speed* of task execution, the *functionality* of the system, the *quality* of the results, the *speed of learning*, the *mental load*, and *user satisfaction*.

Table 1 lists some measurements obtained in our experiments. They show that using only mouse or touch screen leads to more system-user interactions than in the case of multimodal input. Especially the number of actions per experiment differs significantly between the monomodal and multimodal interfaces. However, more interactions do not automatically lead to higher success in image retrieval.

The user feedback that was gathered from the questionnaires (see Fig. 6) corroborates these findings. As seen in Fig. 6(e) the *touch screen and speech* group rated their interaction to be most efficient. Even though the speech recognition component is more error prone and using speech needs some preparation and exercise the multimodal input devices causes slightly less anger than the monomodal ones (see Fig. 6(d)). Furthermore, relying on the mouse for retrieval not only causes slightly more anger, but our subjects also felt that it was unnecessarily complicated. Fig. 6(a) and Fig. 6(b) illustrate that multimodal interaction appears to be more interesting and easier to learn than mouse or touch screen usage. In conclusion, our subjects well appreciated and accepted multi-



(a) is interesting (b) is easy to learn (c) is complicated (d) causes anger (e) is efficient

Fig. 6. Results of a usability questionnaire. For each interaction modality, each aspect had to be rated from 1 (no) to 5 (yes).

Modality	S_E	$T_E = \frac{\text{time[s]}}{\text{experiment}}$	$A_E = \frac{\text{\#actions}}{\text{experiment}}$	$FB_E = \frac{\text{\#feedbacks}}{\text{experiment}}$	$N_I = \frac{\text{\#iterations}}{\text{experiment}}$	$T_I = \frac{\text{time[s]}}{\text{iteration}}$	$A_I = \frac{\text{\#actions}}{\text{iteration}}$	$FB_I = \frac{\text{\#feedbacks}}{\text{iteration}}$	$FB_I^{\text{positive}} = \frac{\text{\#pos feedbacks}}{\text{iteration}}$	$FB_I^{\text{negative}} = \frac{\text{\#neg feedbacks}}{\text{iteration}}$
M	55.6%	129.96	58.81	11.44	4.67	32.10	13.39	2.90	3.01	1.63
T	54.5%	128.94	50.24	17.03	4.45	29.58	11.41	3.93	4.27	3.41
MS	43.3%	131.87	39.30	9.97	3.73	33.13	10.92	2.44	3.39	2.00
TS	66.7%	120.97	29.00	7.23	3.43	35.24	8.20	2.07	3.86	0.65

Table 1. Summary of averaged experimental results with regard to input modalities.

modal image retrieval as offered by our system.

5. CONCLUSION

In this paper, we proposed and evaluated an approach to interactive image retrieval for application in smart environments. Based on the natural modalities speech and gestures for interaction with super sized, wall-mounted touch screens, our CBIR system allows for easy and intuitive image retrieval. An extended usability study underlines that multimodal image retrieval not only yields high user acceptance and easy database access, but also enables successful content-based retrieval. It therefore appears to be a possible approach for browsing digital photo libraries and sharing visual memories in tomorrow's home environments.

6. REFERENCES

[1] M.H. Coen, "Design Principles for Intelligent Environments," in *Proc. of AAAI*, 1998, pp. 547–554.

[2] S. Shafner, "The New EasyLiving Project at Microsoft Research," in *Proc. of DARPA/NIST Smart Space Workshop*, 1998, pp. 127–130.

[3] B. Eggen, G. Hollemans, and R. v.d. Sluis, "Exploring and Enhancing the Home Experience," *Cognition, Technology and Work*, vol. 5, no. 1, pp. 44–54, 2003.

[4] T. Kämpfe, T. Käster, M. Pfeiffer, H. Ritter, and G. Sagerer, "INDI – Intelligent Database Navigation by Interactive and Intuitive Content-Based Image Retrieval," in *Proc. ICIP*, 2002, vol. III, pp. 921–924.

[5] T. Käster, M. Pfeiffer, C. Bauckhage, and G. Sagerer, "Combining Speech and Haptics for Intuitive and Efficient Navigation through Image Databases," in *Proc. ICMI*, 2003, pp. 180–187.

[6] Y. Rui and T.S. Huang, "Optimizing Learning in Image Retrieval," in *Proc. CVPR*, 2000, pp. 236–243.

[7] X.S. Zhou, Y. Rui, and T.S. Huang, *Exploration of Visual Data*, Kluwer Academic, 2003.

[8] G.A. Fink, "Developing HMM-based Recognizers with ESMERALDA," 1999, vol. 1692 of *Lecture Notes in Artificial Intelligence*, pp. 229–234, Springer.

[9] S. Wachsmuth, G.A. Fink, and G. Sagerer, "Integration of Parsing and Incremental Speech Recognition," in *Proc. EUSIPCO.*, 1998, vol. 1, pp. 371–375.

[10] S. Wachsmuth and G. Sagerer, "Bayesian Networks for Speech and Image Integration," in *Proc. AAAI*, 2002, pp. 300–306.

[11] J. Preece, Y. Rogers, and H.C. Sharp, *Beyond Human-Computer Interaction*, Wiley & Sons, 2002.