Empirical Studies in Multimedia Retrieval Evaluation

Mathias Lux[1], Gisela Dösinger[2], Günter Beham[2]

[1] ITEC
Klagenfurt University
Universitätsstrasse 65-67
9020 Klagenfurt
[2] Know-Center
Inffeldgasse 21a
8020 Graz
mlux@itec.uni-klu.ac.at
gdoes@know-center.at
gbeham@know-center.at

Abstract: The evaluation of retrieval mechanisms for inter-method comparison is necessary in academic as well as in applied research. A major issue in every evaluation is in which way and to what extent the actual perception of the user from the target user group is integrated. Within multimedia retrieval systems the impressions and perceptions of users vary much more than in text retrieval. Empirical studies are a common tool in social science and offer a way to research the correlation between the user perception and the computed similarity between pairs of multimedia documents or a query and the set of results. This approach can be used to complement and extend current evaluation approaches. Within this contribution we summarize general methods from social science and psychology for the interested reader in the area of computer science with some knowledge about statistics. Furthermore we give two examples of undertaken empirical experiments and their outcomes. Within the first one the perception of users is investigated and compared to factors like background and gender, while in the second study metrics are tested upon their ability to reflect the notion of similarity of users. Both experiments aim to give examples and insight on how empirical studies can be used in multimedia research in general and multimedia retrieval evaluation in special.

1 Introduction

In computer science, which derives a lot of research methods and traditions from mathematics, it is generally assumed that computers operate on logical levels, where true and false can be distinguished clearly. Research areas and trends like fuzzy computing or genetic programming already drift away conceptually from this assumption. As soon as user interaction like searching documents or creating content becomes part of a process, true and false become concepts which might not be appropriate any more for the situation. An even more complicated issue is to deal with information that algorithms and programs cannot understand and interpret (yet), like video streams or digital photos.

In multimedia retrieval many different retrieval methods have been developed over the last years. The importance of multimedia retrieval has for instance been identified in [Al03], but they also identify the problem that multimedia retrieval evaluation is a far more challenging task than the evaluation of text retrieval mechanisms. One major problem is for instance that there are by far not as generally agreed and as comprehensive standardized test data sets available as in text retrieval.

The evaluation methods from the area of text retrieval were adopted for multimedia retrieval by for instance TRECVID, CLEF (with imageCLEF) or INEX Multimedia, which are prominent retrieval evaluation initiatives. At this point one can ask why not to use other evaluation methods in multimedia retrieval than the ones, which are used in text retrieval. One major problem is that there are currently no other generally agreed methods in the area of multimedia retrieval research. A promising candidate for the evaluation of retrieval mechanisms with the integration of user perception is the empirical study. The area of empirical studies has a long history in social sciences, but especially in the last years there were discussions about the relevance and the scientific value of findings based on empirical studies, and the usefulness and the possible flaws of significance tests.

Empirical studies cannot yield fully true (or false) conclusions. But with the tool of empirical studies conclusions at certain probability levels can be drawn. Used carefully it is a powerful tool even for computer scientists. This is especially the case when a distinction between true and false could not be made anyway, as for example in multimedia retrieval evaluation.

The contribution is structured as follows: The introduction is followed by a brief section introducing related aspects and work. After the introductory words a literature survey on empirical studies is given. After this theoretical part two different example studies are presented. The first one aims at discovering insights on the perception of different (user) groups and the second aims at the evaluation of metrics for the retrieval of digital photos. Finally, this contribution is concluded and directions for future work are given.

2 Related Work

Evaluation of multimedia retrieval algorithms generally relies, like in text retrieval, on huge datasets containing sample queries (called *topics*) and respective optimal result lists. Retrieval algorithms and search engines, that are meant to compete with each other, are bound to incorporate the test data and relate their results gained from the sample queries with the optimal result lists. Measures like *relevance* and *recall* (see for instance [BR99] or [Ri78]) allow then a comparison of retrieval methods. Semantic aspects of multimedia retrieval above the level of keywords and genre classification often remain unmentioned. In *TRECVID*[1], a forum for the evaluation of video retrieval algorithms, for instance focus is put on shot detection, high level feature extraction and search based on topics. The topic based approach as well as the high level feature extraction are also used by the *Cross Language Evaluation Forum* (CLEF) in the imageCLEF[2] track (another evaluation forum) and the *Initiative for the Evaluation of XML Retrieva*l (INEX) multimedia track[3].

In the field of human computer interaction (HCI) statistical methods are a common tool. Many research groups also use statistical methods for the evaluation of user perception and the accuracy of algorithms compared to the view of the actual user. Many of them, like for instance Rodden et al. in [Ro01], where the practicability of content based image organization was surveyed, or Tilinger and Sik-Lanyi [TS06], where the difference between left handed and right handed users in navigation in 3D environments is evaluated, just interpret the results of questionnaires and captured samples subjectively. They do not systematically formulate hypotheses and research questions. A better example is the publication of Andrews et al. [An02], where test setting and samples were described and the significance of the found results was discussed. Even so the study did not match all constraints of an empirical study: No hypothesis was formulated and for result interpretation a mixed approach based on qualitative and quantitative observations is used. Although perfectly matching the requirements of common HCI evaluations the actual method can be further enhanced to match the standards of empirical studies.

The actual perception of the user and the instrument of human vision have been subject to many publications. In [GW01] and [Pr01] the physiology of the eye and the actual physiological abilities of human visual perception are described. They relate the human vision to compression methods and the importance and accuracy - in matching the visual abilities of humans - of colour spaces. The actual notion of image semantics - the meaning behind the pixels - is not topic in those publications. Furthermore in [GC06] several approaches to interrelate human perception to Quality of Service (QoS) aspects are presented.

Within the ITU recommendation ITU-R BT.500-11 “Methodology for the subjective assessment of the quality of television pictures” (see [ITU02]) methods for the evaluation of the user notion of the quality of television pictures are described. The recommendation includes instructions on how the actual testing and analysis (including calculation of significance of found medians) are planned and undertaken. Therefore it provides the reliability of a well planned empirical study in the domain of television image quality assessment. Although the presented approach is very well described the aimed application is different to the evaluation of multimedia retrieval mechanisms. Furthermore the provided instructions in the recommendation are not generic enough to adopt them to other fields.

3 Methodological approach

As the area of statistics and empirical studies is very broad, we can only highlight most important terms and steps to be considered when planning and conducting empirical studies and significance tests. The introduction is mainly based on work and findings presented in [BD02], [Bo99], [CBA84], [FN92], [Hu05], [RJ00] and [Wi99], where the interested reader can find additional and detailed information. We consider an introduction to empirical research in connection to computer sciences important, since usually empirical research is more a matter of social science than technical science education.

---

[1] http://www-nlpir.nist.gov/projects/trecvid/
[2] http://ir.shef.ac.uk/imageclef/2006/
[3] http://inex.is.informatik.uni-duisburg.de/2006/mmtrack.html

3.1 Foundations of empirical studies

The actual core of any empirical study are *research questions*, which define what the study actually investigates. From the research questions so called *empirical hypotheses* are derived, which are preliminary answers to the research questions. They specify expectations concerning certain facts.

Empirical hypotheses are translated into *statistical hypotheses*, which represent them in the form of statistical units and their value. Hypotheses can be differentiated into null hypotheses H_0 and alternative hypotheses H_A where null hypotheses are the ones intended to be rejected in favour of the alternative hypotheses. The aim is to statistically reject the null hypothesis and therefore support the empirical hypothesis. Note that the empirical hypotheses cannot be proven but only retained with statistical means. To give an example, an empirical hypothesis H_0 could say that men and women differ in average reaction time for deciding about picture similarity. The statistical hypothesis H_0 would be $\mu_1-\mu_2=0$, stating that the samples are drawn from populations whose parameters μ_1 and μ_2 are identical.

Concerning the range of hypotheses we distinguish singular, existential, and universal hypotheses:

Hypothesis range	**Hypothesis holds for …**
singular	on single subject of the population
existential	At least one subject of the population
universal	every single subject of the population

Usually quasi universal hypotheses are used, which refer to a restricted population. Depending on the kind of expectations it is differentiated between correlation, differential, and change hypotheses, which might be directional or non-directional. Correlation hypotheses state a covariation between variables, differential hypotheses state that groups of subjects differ regarding a certain variable, and change hypotheses state the changing of a variable over time. Directional hypotheses state that for example reaction time is shorter for men than for women, whereas a non-directional hypothesis states that there is a difference, not specifying the direction.

A hypothesis qualified for research has to fulfil certain criteria. It has to be

- consistent – hypotheses must not contain contradictions in their logical assembly
- criticisable – it has to allow for observations, which might falsify the hypothesis, and
- operationalisable – it has to be assured that the terms used in the hypothesis can be assigned observable phenomena, namely empirically observable indicators.

There are different kinds of empirical studies: it is differentiated between laboratory and field investigations, and experiments and quasi experiments. In a laboratory investigation confounds, which are interfering variables, related to the investigation itself such as noisiness are controlled, which is not the case in a field investigation. In an experiment confounds related to the subjects such as motivation are controlled, which is not the case in a quasi experiment. In any kind of investigation at least one variable, called independent variable, is systematically varied.

Example: Findings in multimedia retrieval might suggest that the pace, at which humans judge visual similarities, depends on the interest in a topic. One possible research question would be: Does the interest in a topic determine the pace at which visual similarity judgements are executed? For a corresponding empirical hypothesis, the directional alternative hypothesis H_A, would say: *There is a positive correlation between interest in a topic and the pace of judging the visual similarity of topic specific pictures*. The respective null hypothesis H_0, which we want to reject, is: *There is no positive correlation or no correlation at all between interest in a topic and the pace of judging the visual similarity of topic specific pictures*. Since we are working with correlations the statistical hypothesis would be $\rho>0$. Here we have universal correlation hypotheses, since a covariation is postulated for all cases and no restriction is made concerning population. These hypotheses also fulfil the criteria of being consistent, criticisable, and operationalisable, since there are ways of determining or assessing interest, judgement pace, and subjective visual similarity.

3.2 Experimental Design & Procedure

The experimental design refers to the logical set up of the study, which allows testing the hypotheses in order to reject or support them.

To undertake an experiment a sample has to be drawn. The sample may consist of human subjects or objects obeying the defined inclusion or exclusion criteria. Different kinds of samples are possible: A *random sample* – each unit of the population has the same chance of being selected – allows for a good generalisation, but especially true random subject samples are very difficult to obtain. *Convenience samples* are more usual, they are drawn at the convenience of the researcher and the availability of subjects or objects. Of course, in this case generalisation of the findings of the study is restricted (see e.g. [Wi99]). Next to the purpose of the study and the population, the sample size is determined by meeting constraints based on the following factors:

- The level of precision, which is the range, within we expect the true value of the population to be located. The sample has to be big enough to guarantee a certain precision.

- The confidence level, which indicates the certainty, the observed value will lie within the range of precision. The sample size has to be large enough to guarantee a certain confidence level.

- The degree of variability, which concerns the distribution of an attribute in the population. The larger the variability, the larger the sample size is required for reaching a given level of precision.

After having drawn the sample, groups of subjects or objects have to be assigned to experimental conditions corresponding to levels or combinations of levels of independent variables. This could be a certain set of stimuli, which are presented to the subjects. Biases are controlled best when subjects or objects are randomly assigned to conditions. If random assignment is not possible so called *confounds* have to be controlled, namely variables, which systematically distort the relationship between independent and dependent variables. Usually one can distinguish between experimental and control groups, where the latter serves as a kind of reference. For example, while an experimental group would undergo a certain treatment, the control group would not undergo it. Depending on the kind and complexity of test arrangement – performing pre-test and post-tests, using one or more groups, repeating measurements, to mention only some aspects – different experimental designs can be distinguished.

For obtaining the required numerical data several techniques and instruments are described in literature, common examples are questionnaires, study of behaviour, or controlled presentation of stimuli. Observations of any kind are translated into data, which might adopt different scale levels. We differ between nominal, ordinal, interval and ratio scale for measurement. The nominal scale allows assertions concerning equality and inequality, the ordinal scale about larger/smaller relations, the interval scale about the equality of differences, and the ratio scale allows assertions about the equality of ratios.

Scale	**Description**
nominal	classification using symbols, e.g. male & female
ordinal	symbols with a pairwise relation, e.g. ordering, for instance strongly disagree, disagree, agree and strongly agree
interval	accurate distance between values can be calculated, e.g. numerical values between 1 and 10.
ratio	a meaningful zero point is available, e.g. time scale or weight scale

Note that various statistical tests require certain scales and data properties such as for instance a normal distribution of the obtained numerical values. In that sense the *kind of data* and the planned *statistical tests* have to be determined a priori.

When the investigation is carried out, it is important to care for standardisation of the used methods to allow comparability and the communication of conditions to the readers to allow reproduction of the study results. Furthermore each subject or object of investigation must find oneself within the experiment in the same situation only differing by experimental condition. Depending on the matter of investigation single or group testing is possible. Experimenter biases also have to be eliminated or controlled: The instruction part for instance, where the subjects of investigation are briefed in advance of the tests, is very important: Subjects must fully understand what is expected from them and they must be introduced to the experiment always in the very same way.

3.3 Evaluating the Data

Certainly a lot can be said about statistical analysis and statistical test theory. The most common and widely used technique is the *t-test*, which is also used in our first example study (see below) to find a significant difference between the mean of two variables. The t-test is used to determine whether a null hypothesis is retained or rejected. It is for instance applied to test expected values based on a population following a normal distribution, with the same variance. The expected values, for instance the mean, are transformed into a test value *t*, which follows a student-t distribution for a valid hypothesis H_0. Using the student-t distribution the probability of validity of H_0 based on *t* can be calculated, which gives the significance level (see e.g. [FN92] and [BD02]). For this introduction we choose to give a brief introduction and summary to the method of correlation tests, to provide a deeper understanding of the second example study.

The contiguity between two variables – correlations can be interpreted as coincidence but not as causality – is statistically represented by the correlation coefficient ρ. The correlation coefficient results from standardising the covariance, which reflects the extent of the linear association between two variables. It can adopt values from [*-1,1*]: *-1* indicates a perfect negative correlation, *1* indicates a perfect positive correlation, and *0* indicates no correlation (see for instance [FN92] or [BS04]). To give an example, a negative correlation between interest in a topic and the pace of judging the visual similarity of topic specific pictures means that a high interest in a topic goes along with a slow reaction time. Depending on the scale levels of the variables different correlation coefficients can be calculated. So for example the point biserial correlation coefficient is based on dichotomous and interval scaled variables, a product moment correlation coefficient is based on two interval scaled variables and a rank correlation, like Spearmans rho, is used for instance for two ordinal scaled variables. For a statistical validation a bivariate normal distribution is required. In this case ρ is a good estimator of the contiguity. If this condition is not fulfilled the estimate is out by $1/n$ but with an increasing n the inaccuracy is negligible. If an empirical correlation proves the null hypothesis, namely whether $\rho=0$, can be decided by a *t*-test. If the calculated *t* is larger than the critical *t*, which can be read from corresponding tables – its size depends on the determined significance level and whether the test is one or two sided – the correlation is significant. To investigate whether two correlation coefficients differ significantly the correlation coefficients are transformed into Fischer z values and z value difference is calculated. The probability corresponding to z can be read from given tables in the literature or can be computed with common statistical software. If it is equal to or smaller than 0.05, representing a significance level of 5%, which is an overall agreed border for significant results, the coefficients differ significantly from each other. If it is equal to or smaller than 0.01 the difference between the coefficients is strongly significant, representing a significance level of 1%, which is an generally agreed border for strongly significant results.

The completion of an empirical study consists of interpreting the results. The interpretation of the data has to be done against the background of the sample, the experimental design, sample size, and used statistical tests. Interpretation also includes pointing at constraints and suggesting for improvements (see also [Wi99] for common flaws and further guidelines).

4 Example Study: Measuring Perception & Understanding of Groups

Past research showed that, on average, differences of cognitive abilities between males and females exist [Ha00]. Males benefit from faster manipulation of visual information in the working memory. Females on the other hand score higher when accessing information in the long term memory and achieve a higher perceptual speed. The purpose of this example study was to gather some insights whether these differences interfere with the perception of similarity between digital photos. Additionally a focus was put on how different amounts of additional textual information describing these images and the association with a technical or non-technical university alter the perception. The rationale behind the distinction of the university background is that people with a more technical background tend to approach and solve problems differently than those with a non-technical background. Surveying these information could also lead to insights of possible perceptional interferences. Summarising, this example study attempts to elaborate on interference of gender and university background on the perception of similarity of digital photos. After stating the background and the purpose of the study, a hypothesis H_A was constructed and reads as follows:

Users perceive the similarity of digital photos without additional information, digital photos with some keywords, and digital photos with textual description, based on gender and association with a technical or non-technical university, differently.

Taking this formulated hypothesis two different groups of variables can be extracted as already stated in section 3.1.

There were three independent variables with two and three levels, respectively:

- Gender, two levels: [1] male, [2] female
- Association with an university, two levels: [1] University of Graz, [2] Graz University of Technology
- Additional information to photos, three levels: [1] pairs of digital photos without additional information, [2] pairs of digital photos with some keywords, [3] pairs of digital photos with textual information

There was one dependent variable:

- Similarity judgements

After formulating the hypothesis and extracting variables to be surveyed, the next steps were to describe for instance which materials will be used, how the group of subjects is selected and how the procedure looks like to conduct the data.

4.1 Materials

The first step was to select pairs of digital photos showing more or less apparent differences in high level features. For this study four different pairs of images conducted from different image databases were assembled. Based on the independent variable "Additional information to photos" all four pairs of images appear three times in the questionnaire, each of them attached with different amounts of additional information. Figure 2 shows one pair of digital photos with keywords attached to them. Each of these pairs was printed on a single sheet of paper and assembled in a folder to allow fast browsing through these pages in a short period of time. The pages were also sorted in a way that no two identical pairs follow each other.

4.2 Subjects

The subjects were 14 female and 14 male students. Half of both groups were at the time of the study students of the University of Graz and the other half at the Graz University of Technology, respectively. Data was collected at different places at the university campus to achieve diversity in the field of study.

4.3 Measures

The following question was asked to assess the similarity of a pair of images: "To what extent do you think these two images are similar?" It was directly followed by a 6-point rating scale ranging from 1 (not similar) to 6 (very similar). This scale forced the subject to decide whether the presented images were similar to some extent or not. Due to the fact that an even number of choices was offered, a neutral answer was not possible.

Figure 1 - Example pair of images used in the questionnaire.

4.4 Procedure

The experimenter visited different places at both university campuses and randomly asked students to participate in the study. Each subject was explained that s/he would be requested to assess the similarity of pairs of images. Questionnaire instructions emphasised that each pair of images had to be assessed in at most 20 seconds. The experimenter took care that pages were turned in time.

4.5 Results & Application

To check the distribution of the assessed data the Kolmogorov-Smirnov test (K-S test) was taken. This test can be used to answer the question whether data is normally distributed or not. Compared to the χ^2-test the KS-test is also suitable for small sets of samples. For all statistical evaluation a significance level of 5% was assumed. Results of the K-S test in table 1 for each pair of digital images (PI) showed that all calculated significance levels are above the critical border of 5% and thus data is normally distributed. Satisfying the precondition of normally distributed data a factor analysis could be calculated.

Table 1 - Results of the K-S Test

	PI 1	**PI 2**	**PI 3**	**PI 4**
Two-tailed significance levels	0.205	0.156	0.158	0.153

The factor analysis was done as a principal component analysis identifying whether the dependent variables extracted from the hypothesis are measuring the declared criteria. Principal component analysis (PCA) is a numerical method reducing the number of dimensions a dataset consists of. Used in statistics it is applied to group related variables together to form more general variables (factors) describing a previously assessed dataset. More detailed information about PCA can be found in [La00]. Results showed the existence of four different factors and not three as assumed previously. Trying to figure out how to describe what these four factors are measuring, no obvious criteria could be found. The third statistical method was aimed to find significant differences within the assessed data sets. The hypothesis states that there exist differences in the perception of similarity between male and female as well as between different university backgrounds. The t-test is a commonly used method in statistics to calculate the significance of characteristic data representing relevant information of the hypothesis. For this study the significance of mean differences (female and male subjects) were calculated. Results of the t-test showed no significant differences between the data sets of gender and data sets of university background except for one pair of images (ρ=0.001, M_{male} = 4.29, M_{female}=5.50). Female students assessed this pair of digital images more similar than male students.

These results lead to the conclusion that the interference of different amounts of information describing images with the perception of similarity could not be proven with the formulated hypothesis, the selected pairs of images and the chosen sample. It is also not possible and valid to state about the interference of gender and different university backgrounds with the perception.

Note that this study has NOT shown that there is no difference in perception. We have just dropped our alternative hypothesis and could not reject our null hypothesis. Although the presented test of the user perception and understanding of image similarity in such a limited way (limited in terms of the sample compared to a global population of users) cannot lead to global understanding of user perception and semantics, it can be applied in communities and domains, were multimedia data is needed. A common example is the retrieval of photos by architectural characteristics. The architects using the system will have their own understanding of similarity of pictures. In a naïve approach photos showing similar buildings from similar angles with similar colours are interpreted as similar. In an architectural use case the age of the building, certain details like windows, doors and ornaments or periods, for example Gothic or Romanesque, are more important features for similarity.

5 Example Study: Finding the most Semantic Metric for Image Retrieval

In general a combination of different low level and high level features is used within an image retrieval system to find images matching user needs. Low level features are mostly numeric values and vectors describing characteristics of an image in a way that is (a) useful and (b) efficient enough for retrieval. Furthermore low level features can be extracted from the image content automatically. Examples for those low level features are colour histograms, dominant colours of images and texture characteristics like regularity or coarseness. High level features on the other hand cannot be extracted without additional manual input (see [SJ98], [Bi99]).

Prominent problems in multimedia retrieval are (i) the selection of appropriate features (see for instance [MZE06]) and (ii) the selection of appropriate metrics for the features (see [SGJ01]). Given for example a design company, where for instance web sites and advertisements are designed, employees search for royalty free images, which contain a certain amount of a colour or have a certain texture. For this use case colour and texture based low level features fit perfectly. A medical research team on the other hand might search for X-Rays containing broken ribs. In this case colour and texture based low level descriptors are intuitively of no use. For high level metadata one can distinguish for instance for text description various weighting schemes like TF*IDF (see [BR99]) and BM-25 (see [RZT04]) or for example different methods for word stemming and disambiguation. Through combination and parameterization of different methods various metrics for high level metadata can be defined. Now the question arises: Which metric fits best?

5.1 Setting up the experiment

In our use case the test data set consisted of 96 photos and associated MPEG-7 based metadata. From the MPEG-7 documents only metadata encoded in the Semantic Description Scheme, which allows a graph based description similar to RDF, was used for comparison of the images (see [Ko03] for MPEG-7 and the Semantic Description Scheme). 5 different types of metrics were used to compare the description graphs pair wise:

- The *Maximum Common Subgraph* metric from [BS98]
- The *Error Correcting Subgraph Isomorphism* metric as described in [BBV01]
- A text based metric using the textual descriptions in description graphs omitting the structural information (cosine coefficient on term vectors using TF*IDF and BM-25 weighting)
- A generalized path index metric based on the vector space retrieval model using the cosine coefficient and TF*IDF and BM-25 weighting as described in [LMG06]
- A suffix tree and path based metric for comparing labelled graphs as introduced in [LMG06]

Including different tested approaches and parameter settings 122 different variants of metrics were tested, whereas many more other variants have been tested but not included in the test documentation. The metrics used in the evaluation as well as the test data set are described in detail in [LG05], [LMG06] and [Lu06].

Based on the aim to find the best fitting metric for a test set of manually annotated photos, following hypothesis H_A was formulated: *There is a strong correlation between the test metric m_i and the perception of the user*. The respective null hypothesis H_0 is: *There is no correlation between the metric m_i and the perception of the user*. To prove the hypothesis H_A standard methods from social science were employed: In general a strong correlation is defined by a correlation coefficient of $\rho \geq 0.5$. Default constraints for a significance test in research are a significance level of 5 % and a statistical power of 80% ($1-\beta = 0.8$). Based on these numbers 22 samples are needed for the significance test according to [BD02]. The number of samples can be further reduced if the effect size is increased. In our experiment we chose a minimum effect size of $\rho = 0.6$ and a used 20 samples. In our case a sample is the similarity values between two images. 20 pairs of images were chosen from the possible 9216 pairs.

5.2 Survey & Data Gathering

To model the perception of the user a group of potential users had to be surveyed. As already mentioned the optimal approach is to select the participants of the survey randomly. In our case we used a convenience sample to reduce the scale of the study. In a first attempt a questionnaire consisting of 14 questions was created. Furthermore for each question a slide was created. Each question allowed 5 different answers ranging from “images are very similar” to “images are not similar at all”. Summaries of the semantic descriptions were shown on the slides beneath the pictures in form of keywords. After a pre test with 14 participants following conclusions for the questionnaire and the slide show were drawn:

- A presentation time of 20 seconds per slide is appropriate for all users to rate the similarity of the images.
- According to participants, who also knew the semantic descriptions, the set of keywords is not appropriate to reflect meaning and content the semantic descriptions, which were the actual input for the similarity metrics.

Based on these findings the slideshow was adapted to show full sentences, which could reflect the actual semantic descriptions better. The questionnaire and slideshow were extended to the number of 20 samples and a second pre test with 13 participants (others than the ones from the first test) has been undertaken. Within this test a sample was identified, where the name had to be disambiguated, as there were two persons with the same name (only surnames were used to ensure anonymity of the people shown on the images) within the test set and many users were uncertain about this fact. After disambiguation of the sample the actual survey with 112 participants (first time participants, who did not participate within pre tests) was undertaken.

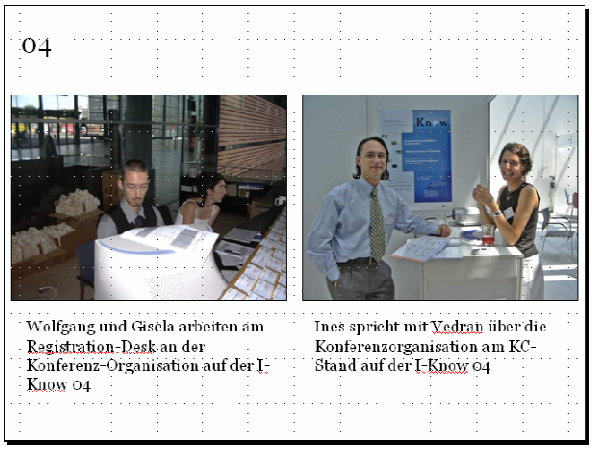

Figure 2 - Example for the slides used in the survey.

For the reference similarity values, which reflect the user perception, the median was taken from the survey results. To ensure the statistical correctness of this approach we also ensured with a χ^2-test that the surveyed data follows a Gaussian distribution with a probability > 99.9 %.

5.3 Correlation Analysis and Results

The first actual significance test was to calculate the correlation coefficient of 20 reference values reflecting the user perception (see 5.2) and the corresponding similarity values of the metrics. Based on the actual correlation coefficient ρ 8 representative metric configurations (weighting schemes, parameters & methods) were chosen. The values of ρ and the 95 % confidence intervals are shown in table 1.

Table 2 - 95% confidence interval

Metric	**ρ**	**left b.**	**right b.**
1. MCS	0.620	0.245	0.834
2. ECSI VS	-0.748	-0,894	-0,457
3. ECSI B	-0.609	-0,828	-0,227
4. VS Text	0.577	0.181	0.812
5. VS BM25 Tripel	0.788	0.531	0.913
6. VS BM25	0.754	0.468	0.897
7. ST IDF Tripel	0.783	0.522	0.910
8. ST IDF	0.791	0.535	0.914

As can be seen from table 2 the confidence intervals of all 8 metrics do not include *0*. Therefore each of the metrics is correlated to the user perception (with 95% confidence). Three of the selected metrics (STF IDF Triple, ST IDF & VS BM 25 Tripel) have a lower border bigger than 0.5, which indicates a rather strong correlation. For all tested metrics the H_0 hypothesis can be rejected at a significance level of 0.01.

Another important question is in how far differences between the correlation values are significant. To answer this question the 95% confidence interval of the correlation difference can be calculated. However with the values shown above in table 1 the correlation difference 95% confidence interval always includes 0, so we cannot assume a significant correlation difference.

5.4 Rank correlation analysis and results

Using the correlation coefficient to investigate a contiguity of two variables also requires the variables to be related linearly. Although the values of the metrics 1, 5, 6, 7 and 8 within the investigated sample follow the same distribution as the reference values (verified with a χ^2-test with a probability > 90%) this cannot be assumed for the metrics 2, 3 and 4 of table 2. In this case a rank correlation coefficient like *Spearmans rank correlation* is a better tool to investigate contiguity between the user perception and the metrics.

Spearmans rank correlation is calculated by converting the actual data into the rank within the sample. Tie ranks are averaged: If for instance 4 samples have the same rank (e.g. 10) they are assigned the median of the respective ranks (e.g. 10+11+12+13=11.5). The actual calculation follows this formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

whereas n denotes the number of samples and d_i denotes the rank difference of the corresponding values of the two investigated variables. The rank correlation analysis has been applied to the same metrics and the same sample – with the intermediate step of ranking the values – as described in 5.3.

Table 3 - Spearmans ρ for the tested metrics and the reference sample reflecting the user perception

Metric	ρ
1. MCS	0.536
2. ECSI VS	-0.721
3. ECSI B.	-0.299
4. VS Text	0.415
5. VS BM25 Triple	0.684
6. VS BM25	0.664
7. ST IDF Tripel	0.658
8. ST IDF	0.720

Based on the table for significance of Spearmans rank correlation coefficient in [Ey04] the H_0 hypothesis can be rejected for the metrics 1, 2 and 5-8 at a significance level of *0.05*. For the metrics 2 and 5-8 this can even be done for a significance level of *0.01*.

5.5 Lessons learned

From the results shown in table 2 we learned in this experiment in a first step that all of the tested metrics are correlated to the user perception. Correcting these results with table 3 - as the use of the correlation coefficient is not expected to provide meaningful results in case of variables with differing distributions - we see that out of the three samples, which did not match the distribution of the reference sample, for two (metric 3 and 4) the H_0 hypothesis could not be rejected. On the other hand metric 2 now also turned out to be a good candidate for multimedia retrieval.

Based on the results we draw the conclusion that the metrics 2 and 5-8 are promising candidates for an image retrieval system based on the users and the images in this use case (see 5.1 and 5.2) as the hypothesis H_0 could be rejected. Although the values of the computed correlation coefficients give a relative ranking of methods, no significant difference between the methods could be supported.

The method applied in the experiment is - although quite laborious and time consuming in the survey & data gathering step - a reasonable alternative to traditional test methods using gold standards and evaluation measures like relevance and recall (see e.g. [BR99] or [Ri79]) as the number of test documents can be reduced. The novelty in this multimedia retrieval evaluation experiment is the usage of standards in empirical research (see [Wi99]), which ensure the possibility of comparison between projects and research groups, and the actual integration of the user group in the evaluation.

6 Conclusion & Future Work

Empirical research is a powerful tool. Keep in mind that a hypothesis can not be proven, but only retained. Furthermore the experiments are labour intensive: In many cases they need pre-experiments and some times do not yield the desired results. Furthermore quantitative studies heavily depend on the questions asked and the way the questions are asked (as already mentioned as experimenter bias in 3.2). Therefore whole chapters of books describe techniques to set up questionnaires and interview scenarios and testing environments.

In our opinion the empirical study has great value for multimedia retrieval: Instead of evaluating independently from users, the target user group can be integrated. Based on observations and heuristics a hypothesis upon user groups, metrics and parameters can be stated and eventually retained based on the results of a study. However the most important point is that the actual users are an integral part in the evaluation. In domain specific search engines, like retrieval of technical drawings, 3D models, sport scenes and more, where subjective similarity differs from domain to domain, it is a common practice that developers, who are no domain experts, plan, implement and adjust the search engines. With empirical studies the accuracy of such search engines can be evaluated in a way, that integrates the actual users, and the subjective similarity can be “captured”.

A next step would be to investigate the accuracy of empirical studies compared to classical retrieval evaluation methods based on test data sets, topics and proposed results. This task can be simplified to evaluate already tested multimedia search engine with empirical studies to allow the comparison of the methods. In addition prepared “standard study templates” should be published, just like the ITU recommendation for the assessment of television quality [ITU02], to ensure a maximum of comparability between different research groups.

References

[Al03] Allan, J. et al. (2003), 'Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, University of Massachusetts Amherst, September 2002', SIGIR Forum 37(1), 31--47.

[An02] Andrews, K.; Kienreich, W.; Sabol, V.; Becker, J.; Droschl, G.; Kappe, F.; Granitzer, M.; Auer, P. & Tochtermann, K. (2002), 'The InfoSky visual explorer: exploiting hierarchical structure and document similarities', Information Visualization 1(3/4), 166--181.

[BBV01] Berretti, S.; del Bimbo, A. & Vicario, E. Efficient Matching and Indexing of Graph Models in Content-Based Retrieval IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23, 1089-1105

[BD02] Bortz J., Döring N., (2002) Forschungsmethoden und Evaluation, Springer Verlag, Berlin.

[Bi99] Del Bimbo, A. (1999), Visual Information Retrieval, Morgan Kaufmann Publishers, San Francisco.

[Bo99] Bortz, J. (1999). Statistik für Human- & Sozialwissenschafter. Springer, Berlin.

[BR99] Baeza-Yates, R. A. & Ribeiro-Neto, B. (1999), Modern Information Retrieval, Addison-Wesley Longman Publishing Co., Inc..

[BS98] Robertson, S.; Zaragoza, H. & Taylor, M. Simple BM25 extension to multiple weighted fields CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management, ACM Press, 2004, 42-49

[BS04] Bronshtein, I.; Semendyayev, K.; Musiol, G. & Muehlig, H. (2004), Handbook of Mathematics, Springer.

[CBA84] Chadwick, B. A.; Bahr, H. M. & Albrecht, S. L. (1984) Social Science Research Methods, Prentice Hall,

[Ey04] Eysenck, M. W. (2004), Psychology: An International Perspective, Psychology Press, chapter Research Methods: Appendices.

[FN92] Frankfort-Nachmias, C. & Nachmias, D. (1992) Research Methods in the Social Sciences St. Martin's Press

[GC06] Ghinea, G. & Chen, S., ed. (2006), Digital Multimedia Perception and Design, Idea Group Publishing.

[GW01] Gonzalez, R. C. & Woods, R. E. (2001), Digital Image Processing, Prentice Hall.

[Ha00] Halpern, D. & LaMay, M. L. (2000). The Smarter Sex: A Critical Review of Sex Differences in Intelligence, Educational Psychology Review 12(2), 229-246.

[Hu05] Huber, O. (2005). Das psychologische Experiment. Eine Einführung. Huber, Bern.

[ITU02] International Telecommunication Union (2002) Methodology for the subjective assessment of the quality of television pictures, Recommendation ITU-R BT.500-11

[Ko03] Kosch, H. Distributed Multimedia Database Technologies CRC Press, 2003

[La00] Lay, David. (2000). Linear Algebra and Its Applications. Addison-Wesley, New York.

[LG05] Lux, M. & Granitzer, A Fast and Simple Path Index Based Retrieval Approach for Graph Based Semantic Descriptions Proceedings of the Second International Workshop on Text-Based Information Retrieval, 2005, 29-44.

[LMG06] Lux, M.; Meyer zu Eissen, S. & Granitzer, M. Graph Retrieval with the Suffix Tree Model Proceedings of the Workshop on Text-Based Information Retrieval TIR 06, 2006.

[Lu06] Lux, M., Semantische Metadaten - Ein Modell zwischen Metadaten und Ontologien, *Graz University of Technology*, 2006

[MZE06] Mitrovic, D.; Zeppelzauer, M. & Eidenberger, H. (2006), 'Analysis of the Data Quality of Audio Features of Environmental Sounds', Journal of Universal Knowledge Management (JUKM) 1(1), 4--17.

[MZE06] Mitrovic, D.; Zeppelzauer, M. & Eidenberger, H. (2006), 'Analysis of the Data Quality of Audio Features of Environmental Sounds', Journal of Universal Knowledge Management (JUKM) 1(1), 4--17.

[Pr01] Pratt, W. K. Digital Image Processing John Wiley & Sons, 2001

[Ri79] van Rijsbergen, C. (1979), Information Retrieval, Butterworths.
[RJ00] Reis, H.T. & Judd, C.M. (Eds.). Handbook of Research Methods in Social and Personality Psychology. Cambridge: Cambridge University Press.
[Ro01] Rodden, K.; Basalaj, W.; Sinclair, D. & Wood, K. (2001), Does organisation by similarity assist image browsing?, in 'CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems', ACM Press, New York, NY, USA, pp. 190--197.
[RZT04] Robertson, S.; Zaragoza, H. & Taylor, M. Simple BM25 extension to multiple weighted fields CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management, ACM Press, 2004, 42-49
[SGJ01] Santini, S.; Gupta, A. & Jain, R. (2001), 'Emergent Semantics through Interaction in Image Databases', IEEE Transactions on Knowledge and Data Engineering 13(3), 337--351.
[Si97] Simon Singh, Fermats Last Theorem, 2002
[SJ98] Santini, S. & Jain, R. (1998),Beyond Query By Example, in 'Multimedia Signal Processing, 1998 IEEE Second Workshop on', IEEE, Redondo Beach, CA, USA, pp. 3-8.
[TS06] Tilinger, A. & Sik-Lanyi, C. (2006), Digital Multimedia Perception and Design, Idea Group Publishing, chapter Issues of Hand Preferences in Computer Presented Information and Virtual Realities, pp. 224-242.
[Wi99] Wilkinson, L. & on Statistical Inference APA Board of Scientific Affairs, T. F. (1999), 'Statistical Methods in Psychology Journals: Guidelines and Explanations', American Psychologist 54(8), 594-604.