Sketch-Based Image Retrieval: Benchmark and Bag-of-Features Descriptors

Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur and Marc Alexa

Abstract—We introduce a benchmark for evaluating the performance of large scale sketch-based image retrieval systems. The necessary data is acquired in a controlled user study where subjects rate how well given sketch/image pairs match. We suggest how to use the data for evaluating the performance of sketch-based image retrieval systems. The benchmark data as well as the large image database are made publicly available for further studies of this type. Furthermore, we develop new descriptors based on the bag-of-features approach and use the benchmark to demonstrate that they significantly outperform other descriptors in the literature.

Index Terms—Image/video retrieval, Image databases, Benchmarking

1 Introduction

For most image databases, browsing as a means of retrieval is impractical, and query based searching is required. Queries are often expressed as keywords (or by other means than the images themselves), requiring the images to be tagged. In view of the ever increasing size of image databases, the assumption of an *appropriate* and *complete* set of tags might be invalid, and content based search techniques become vital.

Different types of content based image queries have been suggested and analyzed: example images; rough, blurry drawings of the desired colors; simple outline sketches; and combinations or extensions thereof [1]. We believe that outline sketches are typically easier and faster to generate than a complete color description of the scene. And they can be generated for arbitrary desired images, while example images may or may not be at hand when searching. In addition, input devices change in favor of sketching as touch-enabled devices become more common. In other words, sketch-based image retrieval (SBIR) is a relevant means of querying large image databases.

Several approaches for SBIR have been suggested. However, to achieve interactive query response, it is impossible to compare the sketch to all images in the database directly. Instead, descriptors are extracted in a pre-process and stored in a data structure for fast access. Very commonly, the descriptors are interpreted as points in a high-dimensional space and finding close matches means searching for nearest neighbors in this space. Moreover, image descriptors can be roughly classified into global vs. local descriptors: global descriptors encode specific features of the whole image that suffice to describe the "gist" of the scene [2], while many local descriptors need to be extracted for a single image, with each descriptor describing only a small spatially localized region of the image [3]. While the use of local descriptors is a common approach in example based image retrieval [4], [5], [6], [7], SBIR systems up to now still employ global descriptors and thus inherit their drawbacks, mainly being not invariant to affine transformations.

An important design feature for any descriptor based retrieval system is that the distance metric in feature space correlates with perceptual similarity. To gauge this perceptual similarity, ground truth information from user studies is needed. Interestingly, Forsyth [8] criticizes the design of many existing image retrieval systems for not meeting real users' needs when they are based on image collections that are comprehensively tagged but are typically unrealistically small.

We design a benchmark for SBIR that is based on a large collection of images, several orders of magnitude closer in size to real data sets than other collections. We also design local descriptors for SBIR. The benchmark allows us to show that they better model humans' perceptual metric between outline sketches and images. More particularly, we report on the following contributions:

- In a controlled user study we gather a dataset of more than 30,000 ratings indicating how well sketch/image pairs match. The analysis of the ratings shows (probably for the first time) that human subjects rate the similarity of sketches and images in a predictable and similar way. Thus, the data defines a benchmark that can be used to evaluate how well the results of an *arbitrary* SBIR system correspond to human expectation. Additionally, we define a set of 100,000 Creative

- *M. Eitz, K. Hildebrand and M. Alexa are with the Department of Electrical Engineering and Computer Science, Technical University Berlin, Germany.*

- *T. Boubekeur is with the Signal and Image Processing Department of Telecom Paris/CNRS, Paris, France.*

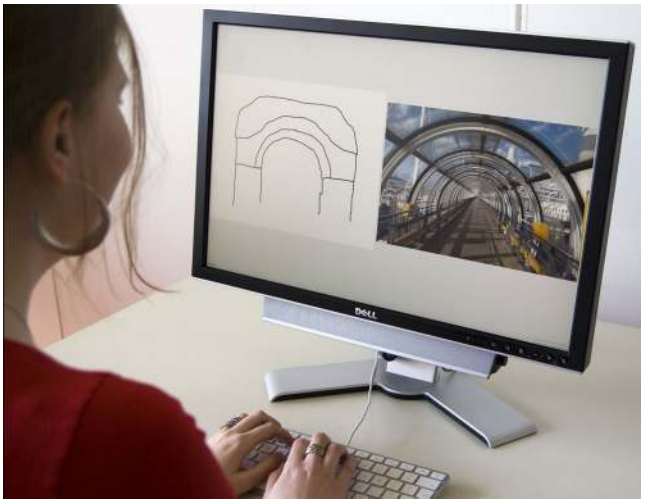

Fig. 1. A display as presented in the user study.

Commons images that is used along with the benchmark. We will make the image database as well as the benchmark freely available for other researchers to compare image retrieval algorithms and evaluate SBIR systems.

- We adapt existing image features (such as shape contexts or SIFT) to SBIR and also introduce new local shape features. These features are used in a bag-of-features approach for SBIR. Based on our benchmark we can show that the bag-of-feature approaches generally outperform existing global descriptors from the literature – we attribute this mostly to translation invariance. And among the bag-of-features approaches the new feature descriptor we designed specifically for SBIR indeed performs better than other general image features.

Based on the new descriptors, we introduce prototypes for several applications that benefit from SBIR.

2 Setup of the Experiment

We assume that humans generally agree that a simple sketch may be perceptually closer to some images than to others. However, we make no assumptions on the similarity of this behavior across humans or even if humans would rate the likeness of images similarly. Given a few examples of similarity ratings, we ask if humans gauge other pairs of sketches and images consistently over time and across subjects. If so, we would like to gather these measurements for comparison with measurements based on feature extraction in computer systems.

To answer this question, we conducted an experiment, in which we show pairs of sketches and images (see Figure 1) under controlled conditions to human subjects and ask them to rate the similarity of the pair. We discuss the following design choices for the study:

- Sketches shown in the study
- Images paired with the sketches
- Human-computer-interaction during the experiment to gather the ratings
- Benchmarking SBIR systems using the results of the study

In the following section, we describe how the experiment was performed and comment on the results.

2.1 Gathering input sketches

The choice of query sketches essentially defines the difficulty of the resulting benchmark. It is important to define it in such a way that current systems can be evaluated. If the sketches would contain too much abstraction, current systems would perform very badly and an evaluation would have to deal with more noise. Overall, we tried to select the set of sketches for the study such that a reasonable compromise between users' demands towards a query system and the capabilities of current systems is achieved.

We gathered a large number of sketches drawn by different subjects that are not biased towards potentially working with a particular SBIR system. In order to achieve a wide spectrum of drawing styles we asked a total of 19 subjects to draw sketches using two different methods: first, 16 individuals (that had not been exposed to SBIR) generated *arbitrary* input sketches depicting arbitrary scenes or objects, drawing them in a way they would expect to work well for an *imaginary* retrieval system. For inspiration we gave them a list of categories (plants, animals, skylines, landscapes, machines, humans) but we stressed that sketches from any other category were also allowed and actually desired. Second, we asked three subjects to also create sketches by roughly tracing the objects seen in existing color images. For both types of sketches, we emphasized that including shading information and cross-hatching techniques should be avoided, instead we asked to create rough outlines that are quick and simple to draw.

This resulted in a total of 164 sketches created for the purpose of this study. From this set we manually selected 49 sketches with the aim that they match reasonably with a sufficient number of images in the database. We used the following criteria in our selection: sketches should depict shape rather than symbols; they should depict non-fictional objects; if any perspective was used it should be reasonably realistic. Using this set of sketches, we now describe how the images for presentation and ranking were chosen.

2.2 Defining sketch/image pairs

Performing the user study requires us to select a certain number of images to be presented for rating along with the sketches. Randomly sampling images from a large image database turned out to not be

applicable: the number of non-relevant images in the database given a certain sketch is large and would likely cause frustration throughout the experiment and also create insufficient data for close matches.

Ideally, we would like to gather a set of sketch/image pairs exhibiting the following properties:

- Rating the complete set should be possible in about one hour per participant in order to avoid fatigue. We settled for 31 sketches (allowing for a wide variety of scenes and objects depicted) and 40 images associated with each sketch.
- The set should generate roughly the same number of ratings for each discrete rank on our ranking scale (i.e. approximately same number of well matching and poorly matching images associated with each sketch). This, however, would require a-priori knowledge of how subjects rate the matches. To approximate this property, we performed a pilot study.

We generated *preliminary* sketch/image pairs by using a subset of the best-performing SBIR systems from Eitz *et al.* [9].

For each of the 49 sketches selected in Sec. 2.1, we queried for the top ranking 20 images using three variants of Tensor descriptor as well as the HoG descriptor (we discuss this process and its potential to generate a biased sample later). This resulted in a collection of $3 \cdot 2 \cdot 20 \cdot 49 = 5880$ sketch/image pairs (including duplicates). Without duplicates, a set of 4532 sketch/image pairs remained.

Using these sketch/image pairs, we performed a pilot-study with three participants, gathering $3 \cdot 4532 = 13596$ ratings. As expected, the distribution of ratings was not uniform for most sketches (see Fig. 7). We now show how to select a subset of these sketch/image pairs that would likely generate a more uniform distribution of the ratings in the final study.

Assuming a uniform distribution of the ratings over the scale 1 to 7, 40 images per sketch would lead to an average of $40/7 \approx 5.71$ ratings per discrete rank on the scale. For each sketch we computed the standard deviation σ of the *number* of ratings from the desired mean of 5.71, and discarded the 18 sketches with the largest σ (exclusively corresponding to sketches associated with many poorly matching images). For the remaining sketches, we randomly sub-sampled the set of associated images such that exactly 40 images remained, however, biasing the sub-sample towards a uniform distribution of ratings. This procedure mostly eliminated images with poor ratings, as expected.

The result of the pilot study is a set of 31 sketches, with exactly 40 images associated with each sketch. This resulting set of 1,240 sketch/image pairs is used in the final user study and presented to the participants for rating. The distribution of ratings from the pilot study and the final study is visualized in Fig. 7 and illustrates clearly, that the proposed subsampling strategy was successful in selecting sketch/image pairs that are closer to a uniform distribution of ratings.

2.3 Interaction for gathering the ratings

We showed each sketch/image pair side by side on a 24 inch LCD monitor under controlled environmental conditions (i.e. similar lighting conditions, no significant audible distraction, 80 cm viewing distance). An example display is shown in Figure 1. Subjects were exposed to the stimulus for 2 seconds. After the stimulus the screen turned black for 1 second. Subjects were asked to rate the similarity on a 7-point Likert scale [10] from 1 (best) to 7 (worst), by pressing the corresponding number key on a keyboard. The rating could be submitted and changed during the total of 2 seconds of stimulus or the 1 second of blank screen, but not afterwards. After the 3 seconds, the next pair was automatically displayed.

The first set of pairs was presented with a suggested scale (see below). After that, the experiment started and pairs were presented in random order. Participants were allowed to pause the study at any time. Apart from the final rating we also recorded the time until the final rating.

2.3.1 Anchoring

To understand if subjects rate sketch/image pairs consistently over time and, if so, to get comparable ratings across subjects it is important to expose subjects to example pairs and a desired rating. Thus, directly prior to the experiment we display a set of 21 sketch/image pairs (3 sketches, with 7 corresponding images each, as shown in Figure 2) together with their corresponding ratings to each participant.

The idea is that human subjects use this information for anchoring, creating a bias towards these ratings for similar matches: if human subjects are able to consistently rank the likeness of sketches and images then anchoring works. If humans are generally unable to consistently assess the matches then the effect of examples would quickly diffuse and subjects would rate the matches randomly or following some mechanism that we are not aware of. We include the set of sketch/image pairs used for anchoring randomly into the later part of the experiment.

If consistent similarity rating based on visual perception was impossible, this would now be easy to detect by analyzing if subjects significantly changed their rating for the examples or if ratings were inconsistent across subjects despite the anchoring.

2.3.2 Choice of stimulus duration

It is clear that the sketch/image pair cannot be shown too short or too long. Our setup leads to disks with a radius of roughly 2cm being projected into the foveal region of the eye. This means a human observer needs

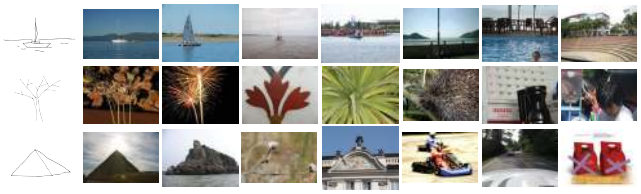

Fig. 2. Sketch/image pairs used for anchoring. Left: sketches; right: associated images ordered according to their rating from 1 (left) to 7 (right).

saccades to compare features in the sketch and the image. Fixations between saccades are rarely shorter than 200ms across different visual activities [11]. In a similar experimental setup and comparable content a recent study reports a mean time of 250ms between saccades [12]. Thus, comparing few features in the pair requires at least a roughly a second, perhaps more. In a preliminary experiment our participants found display times less than two seconds exhausting. On the other hand, more time leads to more high-level considerations and more noisy data. We therefore settled for 2 seconds presentation time. This also helped to keep the time for rating the set of 1,240 sketch/image displays to roughly an hour, which we considered tolerable.

3 Evaluation

Given the experiment we now discuss how to evaluate the results. We propose to base the evaluation on *rank correlation* and show how this approach can also be used to compare automatic SBIR systems against the human data gathered in the experiment.

3.1 Rank correlation

Let x be a list of n ratings with x_i denoting the rating assigned to the i^{th} element. Kendall's tau [13], [14] is a widely used measure of rank correlation, allowing to assess the degree of correspondence between *two* ordered lists of the same elements and determine the significance of this correspondence. In our setup, these lists are generated in two areas: first, participants of the user study rank the benchmark images with respect to a benchmark sketch (visualized in Fig. 3). Second, SBIR systems rank the results of a query with respect to a query sketch. Contrary to linear correlation, as used in a benchmark related to ours [15], Kendall's rank correlation coefficient is independent of the scale and distribution of the data to be compared. Therefore, it facilitates a direct comparison of the experimental data (on a scale from 1-7) with the results of any retrieval system, given the particular system is able to generate a weak ordering of the result images when querying with a particular sketch.

Kendall's rank correlation coefficient is computed as the difference between the number of concordant (similarly ordered) and discordant (reversely ordered) pairs of ranks (x_i, y_i) and (x_j, y_j) in two ordered sets x and y. A pair is concordant if $(x_i - x_j)(y_i - y_j) > 0$ and discordant if $(x_i - x_j)(y_i - y_j) < 0$. Normalization by the total number of pairs is applied to gain independence of the test set size. Let n_c denote the number of concordant and n_d denote the number of discordant pairs then τ is defined as:

$$\tau = \frac{n_c - n_d}{n(n-1)/2} \tag{1}$$

According to this definition, τ can take values in the range $[-1, 1]$, with -1 indicating a reversed list, 0 indicating that the two lists are independent and 1 indicating that the two lists have the same (weak) order.

To summarize, we suggest to use Kendall's tau as a measure to compare the performance of a SBIR system against the data gathered in the user study, thus defining a general benchmark for SBIR systems.

3.1.1 Tied ranks

It is common in our setting that the ranks of pairs in a list are tied, i.e. have the same rank: subjects evaluated 40 images on a discrete scale from 1 to 7, necessarily resulting in a certain number of ties. Also, a SBIR system may produce exactly the same score for two images in the collection, thus possibly producing tied pairs. Our measure of correlation needs to be able to deal with those cases; the denominator in Equation 1 is too large in the case of ties and thus needs to be adapted to keep the correlation coefficient in the range $[-1, 1]$.

Let $N = n(n-1)/2$ (number of possible pairs in a set of n distinct elements), $U = \frac{1}{2}\sum_{i=1}^{t} t_i(t_i - 1)/2$ (number of ties in the first list) and $V = \frac{1}{2}\sum_{i=1}^{u} u_i(u_i - 1)/2$ (number of ties in the second list). Kendall's rank correlation coefficient adapted to the case of ties is denoted as τ_b and defined as [14]:

$$\tau_b = \frac{n_c - n_d}{[(N-U)(N-V)]^{\frac{1}{2}}} \tag{2}$$

3.1.2 Measuring statistical significance

Assuming the null hypothesis of no correlation $H : \tau = 0$ is true, we are interested in the probability of obtaining a correlation coefficient τ greater or equal than the actually observed correlation coefficient τ_o by chance (p-value). In other words: each observed rank correlation value τ_o comes with a corresponding p-value, with lower p-values indicating that it is less likely to observe $\tau \geq \tau_o$ in case the two ordered lists are actually not correlated.

Kendall's tau allows for a very convenient assessment of the significance of a correlation value, since the distribution of correlation value frequencies quickly tends to a Gaussian distribution with

$\sigma^2 = (4n + 10)/9n(n - 1)$ for $n > 10$. For smaller sample sizes, the exact distribution can instead be easily computed. In the case of ties, the distribution also tends to a Gaussian, however with a different standard deviation [14].

Throughout the paper we assess significance of correlation coefficients using a significance threshold of $\alpha = 0.05$, i.e. accepting the null-hypothesis of no correlation for $p \geq \alpha$. For $p < \alpha$ we reject the null-hypothesis and instead accept the hypothesis that the two ordered lists indeed correlate.

3.2 Benchmarking SBIR systems

We propose to use the 31 ordered lists (one per sketch) resulting from averaging the ratings over the 28 study participants as "ground-truth". Each of these lists can be seen as a consensus between the participants and is optimal in the sense that it maximizes the average Spearman's rho rank correlation coefficient between the 28 lists from the data and the estimated one [14].

We use the resulting 31 benchmark lists to evaluate other SBIR system's results by determining Kendall's rank correlation coefficient τ_b for each of the corresponding 31 ordered lists generated by the system. Benchmarking a system thus results in 31 correlation coefficients that for each sketch indicate how well the system ranks the benchmark images compared to human judgement. Throughout this paper, we report the average of those 31 correlation coefficients as the performance indicator for a specific system.

The complete set of sketches and images selected for the benchmark is shown in the additional material. The images corresponding to the sketches are sorted according to the order defined by the benchmark, i.e. a retrieval system achieves high correlation values if it returns the benchmark images in the same order as shown in the additional material.

To summarize, benchmarking an arbitrary SBIR is now straightforward and requires the following steps:

- Place the 1,240 benchmark images in the collection of 100,000 images (note that this is not strictly necessary when using global descriptors)
- For each of the 31 benchmark sketches, perform a query on the resulting collection of 101,240 images and determine the ranking of the corresponding benchmark images in the result set
- For each of the resulting 31 ordered lists, compute τ_b against the corresponding "ground-truth" list.

The resulting values can be compared to the rank correlation coefficient across subjects in the study (see Figure 5). An ideal system would perform similar to human subjects.

3.3 Study and analysis

We gathered our data using the experimental setup described in Sec. 2.3. Our 28 participants had an average age of 25.5 years (± 3.1), 23 were male, 5 female.

3.3.1 Consistency among sketch/image pairs

To understand if the 7 point scale we used in the experiment is not introducing noise, we analyzed the distribution of ratings by grouping the ratings according to the median of the 28 ratings for each sketch/image pair $p_i, i \in \{1, \ldots, 1240\}$. Let $r_{i,j}$ denote the rating of p_i by participant j ($j \in \{1, \ldots, 28\}$). We first compute the median rating m_i for each p_i as $m_i = \text{median}_j(r_{i,j})$. We then define the multiset of ratings that fall into a common bin k when binning according to their associated median rating m_i as

$$\mathcal{B}_k = \{r_{i,j} | k \leq m_i < k + 1\},\ k \in \{1, \ldots, 7\}. \quad (3)$$

For each $\mathcal{B}_k$ we show the mean and standard deviation in Figure 8 (left) and the histogram in Figure 8 (right). The analysis shows that the number of ratings is roughly uniform over all 7 bins with a slight peak at bin 6. The variance of the ratings is smallest for the lower and higher bins and peaks at bin 4 with $\sigma = 1.4692$, i.e. participants were quite consistent for good and poor matching pairs and slightly less consistent for the average matches.

3.3.2 Correlation analysis

We have performed a correlation analysis of the rating results of this study. As discussed we use Kendall's rank correlation coefficient τ_b.

First, we analyzed the correlation of ratings throughout the experiment for the sketch/image pairs used for anchoring. 77.4% of the resulting correlation coefficients can be accepted as significant, the resulting distribution of values is shown in Figure 4. We conclude from this analysis that for most subjects anchoring is significant, i.e. we can accept the hypothesis that the ratings given by the participants are correlated with the ratings shown in the anchoring process.

Second, we compared the ratings *across* subjects. For each sketch, we computed τ_b for all possible pairs of lists generated by the 28 subjects. This resulted in a total of 378 correlation values per sketch – the number of possible pairs for 28 participants. The analysis reveals that the ratings across subjects are mostly consistent, 86% of the inter-user correlation values lie over the significance threshold for $\alpha = 0.05$.

While τ_b describes the correlation between *two* sets, we are also interested in a measure of agreement among the 28 participants (per sketch) as a group. This is known as the problem of m ordered lists and formalized by Kendall's W [14]. We have computed Kendall's W for all 31 sketches – the results show that for most sketches, the agreement between the study participants is strong, with all correlation values being highly significant.

Altogether, we take this as a strong indication that humans agree on how well a sketch fits a natural image – a further indication that sketches can be

Fig. 3. Examples for ordered lists leading to various τ values. First row: benchmark, a subset of the images corresponding to the sketch on the left ranked as defined by our benchmark. Second and third row: different orderings of the same images as in the first row leading to varying τ values.

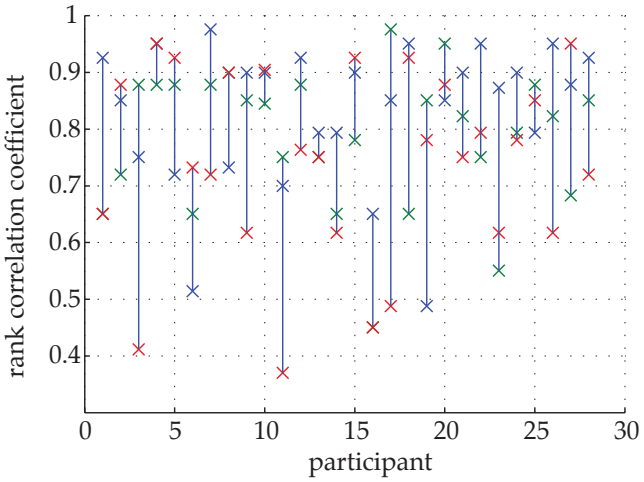

Fig. 4. For each subject, the plot shows three correlations. These values denote how well the subject's ratings correlate with the ratings defined in the anchoring process for the three sketches shown in Figure 2.

used in image retrieval systems and an additional motivation to develop systems that match humans' expectations. This also indicates that a benchmark created from this data is sound and general enough to cover the similarity perception of a large body of different users. For each sketch, we visualize the distribution of τ_b and the value of W in Figure 5.

Third, we analyzed how human subjects perform against the benchmark. Note that in this case Kendall's W is not appropriate since we are indeed only interested in the correspondence between *two* ordered lists: the benchmark's and a subject's. For each subject we therefore computed 31 τ_b coefficients (one for each sketch), estimating the correspondence between the subject's data and the the benchmark. The analysis shows that 99% of the correlation coefficients lie over the significance threshold for $\alpha = 0.05$. We take this as a strong indication that the proposed benchmark adequately reflects human assessment of shape/image similarity. We plot the resulting distribution of τ_b coefficients in Fig. 6.

4 Descriptors & Indexing

We propose using a bag-of-features approach [5] for SBIR employing small local descriptors. This allows basing the search on a standard inverted index datastructure from text-retrieval [16]. We discuss the four main components of our retrieval system: a) definition and representation of local features, b) sampling strategies defining the coordinates in image space of the features to be extracted, c) codebook definition and d) the actual search algorithm based on an inverted index. While we rely on existing methods proposed in the literature for sampling, learning a codebook and defining the search algorithm, our work focuses on feature representation, as this is the part where SBIR systems differ strongly from classical example-based retrieval systems. In particular, we evaluate which feature representations are best suited for implementing a bag-of-features based SBIR system by searching the parameter space for well performing combinations. We base our performance evaluation on the benchmark defined in Section 3, i.e. on real user's preferences.

4.1 Bag-of-features model for image retrieval

Using inverted indices for image retrieval has been inspired by techniques from text retrieval [4] and indeed bears many similarities with text retrieval approaches. However, there are several main differences:

- Text documents are naturally separated into atomic, semantically meaningful entities (words), pixels alone however do not contain sufficient information. Instead, larger image patches are considered as a whole and a feature encoding

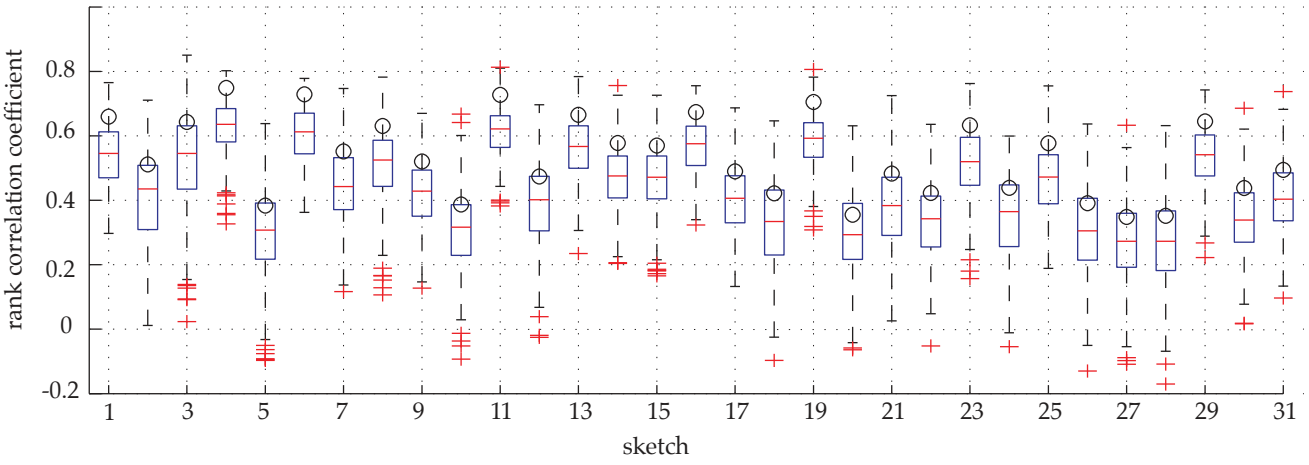

Fig. 5. Distribution of τ_b correlation coefficients between the 28 study participants (per sketch). Median correlation is depicted by a horizontal line inside a box, boxes span the data lying inside the 75th/25th percentiles. Whiskers extend to data points within 1.5 times the interquartile range, other data points are considered outliers and denoted by a (red) plus. Black circles represent Kendall's W coefficient indicating the degree of correlation between the 28 participants taken as a group.

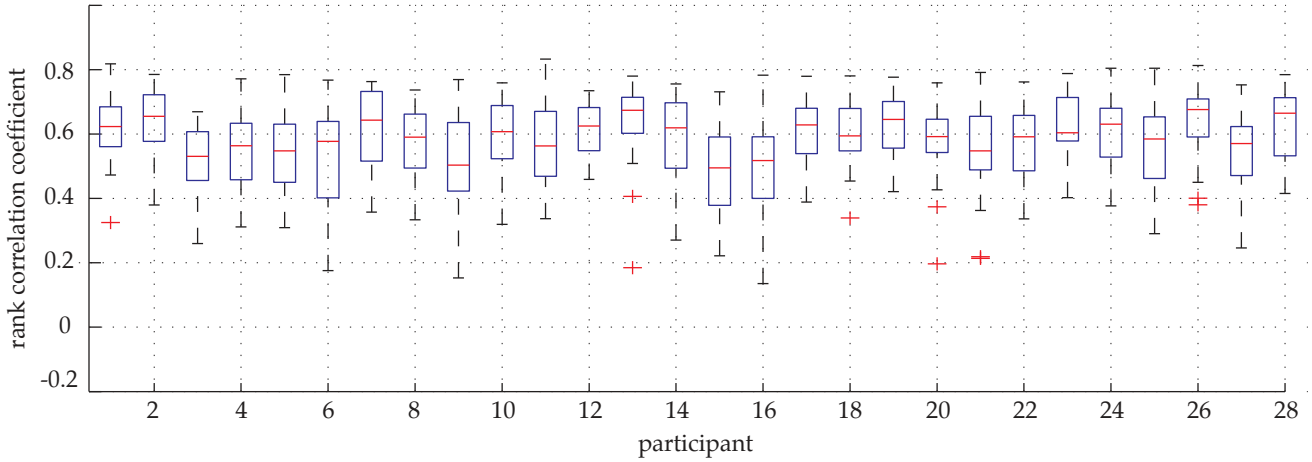

Fig. 6. Distribution of τ_b values between 31 human data lists and the benchmark lists (per participant).

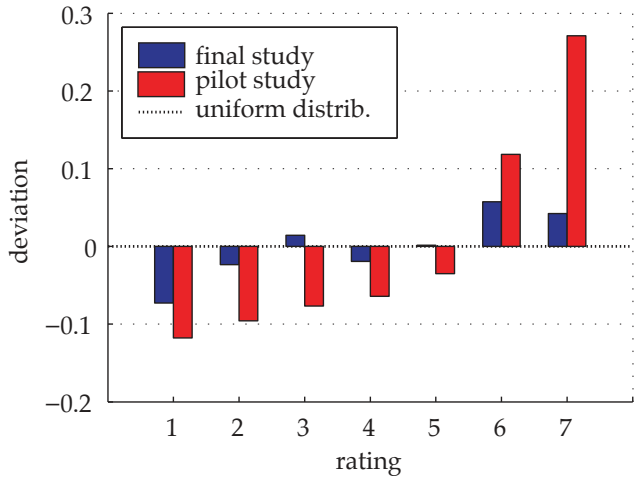

Fig. 7. Distribution of ratings gathered in pilot/final study relative to the desired uniform distribution.

the essential information in this local area is extracted.

- No natural boundary between image patches typically exists, instead the locations of patches in image space need to be sampled.
- Contrary to text retrieval (fixed number of distinct words) a gigantic number of different features could be extracted from images. Therefore a finite sized visual codebook of visual words is typically generated which assigns perceptually similar features to the same visual word.

We next quickly discuss the design choices made in our system for addressing those issues and then discuss the proposed feature representations that we have identified as suitable for *sketch-based* image retrieval in more detail.

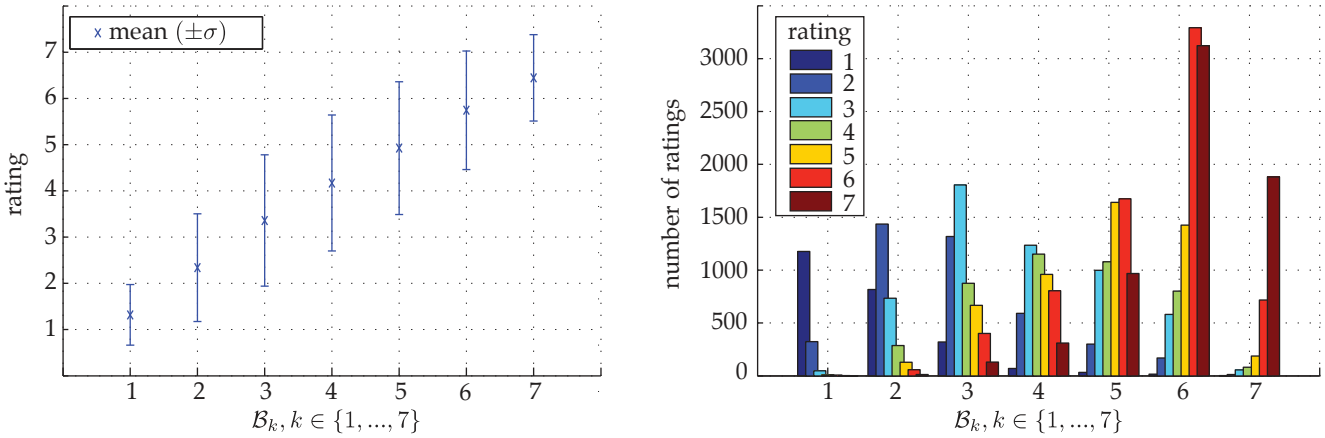

Fig. 8. Distribution of ratings in bins $\mathcal{B}_k$ (see Eqn. 3). Left: mean and standard deviation σ, right: histogram.

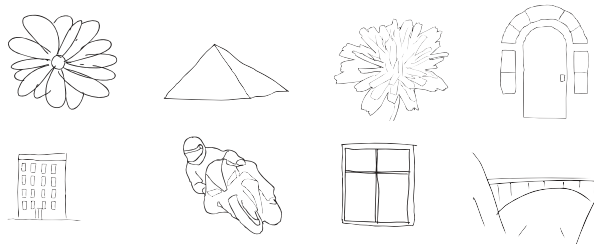

Fig. 9. Best/worst sketches regarding user/user correlation. Best sketches are shown in the top row and correspond to sketches 4, 6, 11 and 19 in Figure 5, worst sketches are shown in the bottom row and correspond to sketches 5, 20, 27 and 28.

4.1.1 Sampling strategies

Initial experiments with the SIFT interest point detector [3] have – as expected – shown that very few to no interest points are detected on our binary sketches. Instead we perform random sampling in image space which has been shown to work well in a bag-of-features setup for image classification [17]. We employ two different strategies, tailored to the features to be extracted from the sample regions: a) generating 500 random samples in image space and b) generating 500 random samples on the sketch lines. In the next section we describe which strategy is appropriate for which descriptor.

4.1.2 Codebook generation

We learn codebooks from a set of 20,000 training images (disjoint from the set of query images). For each training image we extract 500 features sampled on random locations and perform k-means clustering on the resulting set of 10 million feature vectors. We evaluate the influence of the codebook size on retrieval performance by testing the following range of codebook sizes: 250, 500, 750 and 1,000.

4.1.3 Search strategy

We rely on standard methods for creating and querying the inverted index [16], [18]. Creating the index in our implementation for 1 million images takes roughly 4 hours, a query for 50 images takes a couple of seconds, depending on the number of distinct visual words in the query. Note that those timings could be easily improved, this however was not the focus of this work.

4.2 Local features for representing sketches

In this section we describe our approach to identifying suitable local descriptors for SBIR, capable of capturing essential properties of local sketch areas. We start with two popular feature representation approaches, that naturally seem to be suitable for representing our binary sketches: a) shape contexts [19] and b) histograms of oriented gradients [3], [20]. We then explore if and how those descriptors need to be modified to work well for SBIR and evaluate a large range of parameter settings against our benchmark to identify the best performing descriptor. We show the results of our evaluation in Figure 11.

When extracting descriptors from photographs we apply the Canny edge detector [21] in a pre-process. We use a (rather large) $\sigma = 5$ in order to supress small lines that are unlikely to be sketched by users. We use a low threshold of 0.05 and a high threshold of 0.2 (both thresholds are taken relative to the minimum/maximum magnitude of gradient remaining after the non-maximum suppression step).

4.2.1 Shape context

Shape context features encode the distribution of sample point locations on the shape relative to each of the other sample points (see Figure 10, left). The distribution is encoded as a log-polar histogram with 5 bins for (logarithmic) distance and 12 bins for angles. Because each shape context encodes information about all the other sample points, it inherently

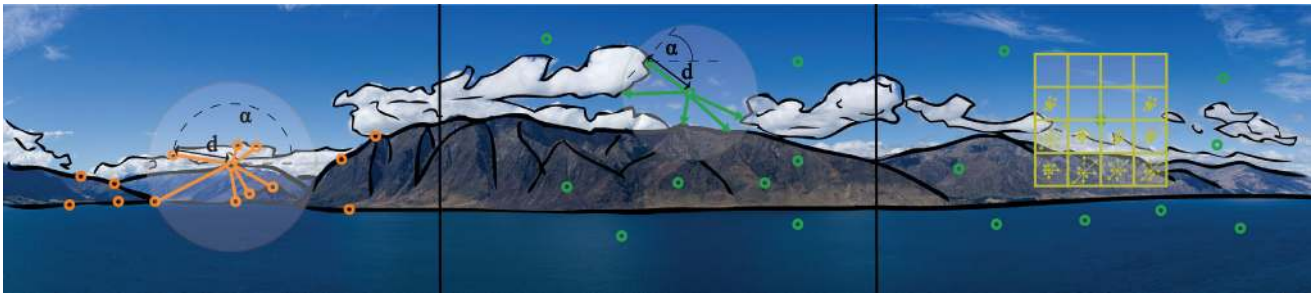

Fig. 10. Left: the (localized) shape context descriptor samples random points on the extracted sketch lines and stores the distribution of the positions of those points relative to its center. Middle: the spark descriptor samples random points in image space, traces rays within a radius until a sketch line is hit and stores the distribution of hit-point parameters relative to its center. Right: the SHoG descriptor employs a histogram of dominant gradient orientation within a local rectangular area around a random sample point in the image.

captures global information about the shape. For our purposes, we *localize* shape contexts by computing the distribution only within *local regions* of the sample point with a fixed radius. We sample 500 points on the feature lines and use several local radii (in percent of the image diagonal): 5, 10, 15, 20, 25, 30. Additionally, we test the global variant, encoding for each sample point the distribution of all other points on the shape. We sample only on the features lines, corresponding to sketched lines for queries and Canny lines in the case of images.

4.2.2 Spark feature

We introduce the *Spark* feature as an extension of shape context features, specialized for the needs of SBIR. While they bear some similarity with shape contexts regarding the information that is encoded, both the way that other sample points on the sketch are generated as well as the local region they describe are different. First, we generate random sample points in the image domain. Note that those sample points must not lie on sketch lines, instead they are generated to lie in the empty areas between feature lines. For each sample point we trace rays of random direction until the first feature line is hit, if any (see Figure 10, middle). The descriptor then stores the distribution of several properties of the feature lines at the hit-points. We have tried several variants, including 2D histograms storing distance/angle information (as in the shape context), distance/orientation information (orientation of the feature line at the hitpoint) and a 3D histogram variant, storing distance, angle and orientation information.

4.2.3 Histogram of oriented gradients

Histograms of oriented gradients (HoG) are usually implemented as 3D histograms encoding the distribution of gradient orientations in small local areas [3], [20]. We test the localized variant used in the SIFT descriptor (4x4 spatial bins and 8 bins for gradient orientation) extracted from a range of differently sized local rectangular windows (in percent of the image diagonal): 5, 10, 15, 20, 25, 30. We determine sampling locations by random sampling in image space.

4.2.4 SHoG feature

We improve the standard histogram of oriented gradients descriptor by storing only the most dominant **s**ketched feature lines in the **h**istogram of **o**riented **g**radients. This helps to make descriptors extracted from user sketches and descriptors extracted from the database image more similar, improving the probability that descriptors that are near in feature space correspond to perceptually good matches. We use Canny lines as an indicator for important features and store only orientations that correspond to data lying under a slightly blurred version of the Canny feature lines. We test the same variants of rectangular windows as with the standard histogram of oriented gradients descriptor. The resulting descriptor is visualized in Figure 10, right.

4.3 Evaluation of descriptor performance

We evaluate performance of the local SBIR descriptors using the benchmark defined in Section 3. We are interested in how the following three parameters influence retrieval performance: a) codebook size, b) local window size for feature extraction and c) feature representation. For each of the six descriptor variants (histogram of oriented gradients, shape context, spark (2D, 2D, 3D), and SHoG) we therefore evaluate all combinations of four different codebook sizes (250, 500, 750, 1,000) and seven local feature window sizes (5, 10, 15, 20, 25, 30, in percent of the image diagonal). We generate the codebooks from a training set of 20,000 images, randomly sampled from 9 million Flickr images which is disjoint from the evaluation set. The evaluation set contains 100,000 images (also randomly sampled from the same set of 9 million Flickr images) plus the additional 1,240 benchmark images. For all combinations, we compute inverted indices using the corresponding codebooks

and the descriptors extracted from the evaluation set of 101,240 images.

We then benchmark all resulting retrieval systems (defined as a combination of feature type, codebook size and local radius) with the method proposed in Section 3, i.e. we perform queries with each of the 31 benchmark sketches and determine the order of the corresponding 40 benchmark images in the result set. We show a visualization of the results in Figure 11 and comment on the findings in the next section.

5 Results and Applications

In this section, we comment on which system parameters mainly influence performance of the proposed local SBIR descriptors and show several experimental applications that benefit from a state-of-the-art SBIR system. We show a demo of the retrieval system as well as all experimental applications in the accompanying video.

5.1 Finding optimal local descriptor parameters

We discuss our findings regarding the influence of codebook size, local feature window radius and feature representation on retrieval performance. Additionally, we present the results of the evaluation of a set of four different *global* sketch-based descriptors: Angular radial partitioning (ARP) [22], edge histogram descriptor from the MPEG-7 standard (EHD) [23] and the Tensor and HoG descriptors [9] and evaluate, whether a bag-of-features based approach is superior to a global approach. All performance numbers have been determined by evaluating the corresponding retrieval system with a certain set of parameters against the benchmark defined in Section 3. A combined plot summarizing the results is shown in Figure 11.

5.1.1 Codebook size

We have generated all codebooks from the same training set of 20,000 images testing a range of different codebook sizes: 250, 500, 750 and 1,000. Retrieval results generally are best for a codebook size in the order of 500-1,000 visual words. Note that this number is significantly smaller than the number of visual words typically used in example based image retrieval. We attribute this to the fact that the information encoded in sketches is much sparser compared to images. Additionally, no information about the magnitude of gradients is available.

5.1.2 Feature size

We find that rather large local feature sizes are necessary to achieve good retrieval results, independent of the feature representation used. If the local window size is chosen too small, the majority of visual words would encode small line segments in a variety of different orientations, which is not discriminative enough to achieve good retrieval results. Note that this result depends on the type of sketch used as input: artists might be able to express fine detail in their drawings and thus might benefit from smaller local windows – all participants that have generated our sketches however have drawn only large scale features. We thus assume that a large majority of potential users would not be able to express fine details in their drawings. This could also explain the good retrieval results of the existing global descriptors. Summarizing, we find that local feature sizes in the order of 20-25% of the image's diagonal perform best. We discuss the implications of this in Section 6.

5.1.3 Feature representation

We find that unmodified existing shape descriptors (shape context [19] and histogram of oriented gradients [3], [20]) are not directly suitable for the task of SBIR. We report the following maximum correlation coefficients achieved for the global descriptors (Tensor and HoG): 0.223. Both the shape context and the histogram of oriented gradients descriptor achieve lower maximum correlation values of 0.161 and 0.175 for the range of parameters tested in this evaluation setup. The proposed spark descriptor attains a maximum correlation coefficient of 0.217 for a codebook size of 1,000 visual words and a local radius of 20. The histogram of dominant local orientations outperforms the other local descriptors with a maximum correlation coefficient of 0.277 for a codebook size of 1,000 visual words and a radius of 25% of the image diagonal.

5.2 Experimental applications

We present several experimental applications that build on the bag-of-features SBIR system proposed in Sec. 4. Note that in each application users are only asked to draw or edit simple binary feature lines.

5.2.1 Specialized search: human faces

We performed initial experiments for sketch-based search on the FERET face database [24], containing 12,000 pictures taken from 1,200 human faces. As illustrated in Figure 13, our SBIR system provides acceptable to good results even when applied in such a specialized search scenario. Controlling the visual search processes through sketch lines opens the question of the drawing interface provided to users – additional computation (i.e. symmetry enforcement, filtering) could help to improve retrieval results in this scenario.

5.2.2 2D search from 3D shapes

We propose a simple strategy for 2D search from 3D objects: the user starts by choosing a 3D object in

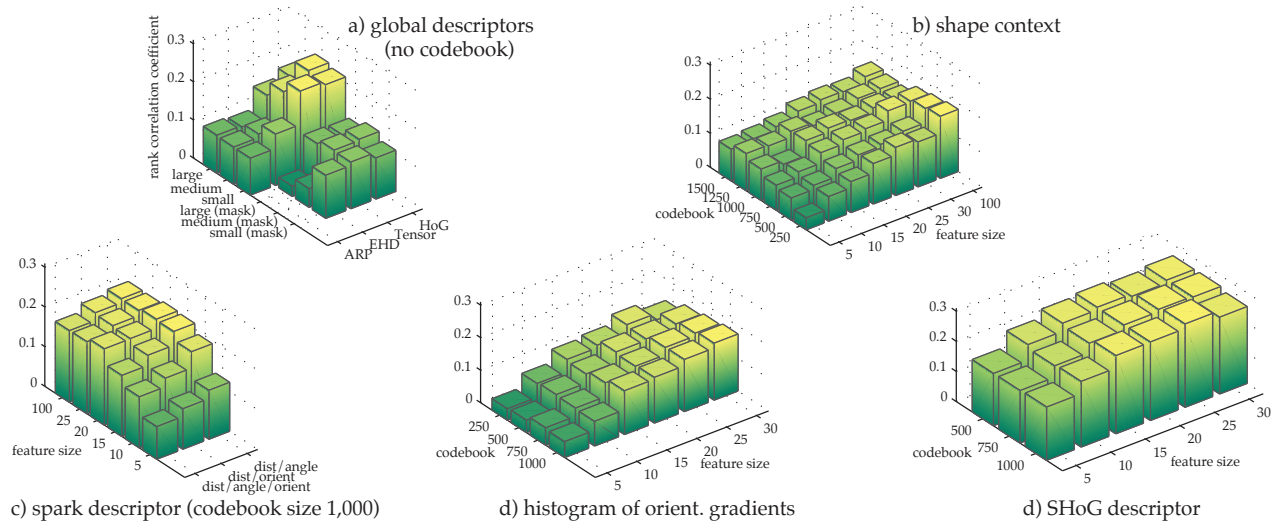

Fig. 11. Evaluation results for global and local descriptors. For all plots we show the evaluation results for varying codebook sizes (250-1,000) and local feature radii (5-30). For the spark descriptor, the codebook size is set to 1,000, shown are the results of the three different histogram types for varying local window sizes. The results of the global descriptors are shown for varying descriptor resolutions and masked/unmasked variants.

Fig. 12. Typical query results of the proposed system (top n images when querying with sketch on left).

Fig. 13. A binary sketch capturing main face feature lines (left) is used to search among thousands of human face photos. Two samples of corresponding results are shown on the right of each sketch.

a shape repository and selects an appropriate viewpoint. We render the view using recent line drawing algorithms [25], [26] and forward to resulting 2D image to our SBIR search engine. While this method obviously requires access to a browsable database of 3D objects, we believe that it could be particularly helpful for 3D artists modeling and texturing 3D shapes: extracting such data from real photos is a frequent task but the search for adequate images is tedious. Our SBIR system definitely eases the search for matching images, by using the information contained in the designed 3D shape to query for matching pictures. Note that an image-based search would not work as the main source of information in the 3D shape is lines and contours. We present an example of such a *2D search from 3D input* in Figure 14.

5.2.3 Basic sketch-based video search

Video collections could also be understood as large image collections, which again can now be easily searched using the proposed SBIR system. We performed initial experiments by searching within a collection of 17 feature-length movies. We extracted the I-frames of all videos, resulting in a set of 78,000 images which again could be searched using the proposed SBIR system. The result were satisfying for

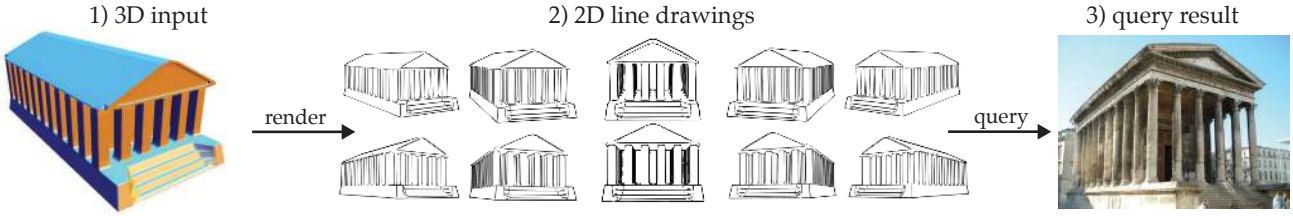

Fig. 14. Image search from 3D input: the image on the right appears among the top 10 results.

simple scenarios but appears that complex animations will require *annotated* sketches, containing additional information on movements [27], [28], [29].

6 Discussion

6.1 User study

There were several design decisions involved in the creation of the user study, all of which influenced the outcome to a certain degree. The choice of sketches and corresponding images certainly has the largest influence: we tried to generate them in such a way that current SBIR systems would be able to achieve reasonable results with the resulting benchmark. Since the matching images were generated from the output of a system using global descriptors, this has some influence on what the benchmark measures: it is not explicitly designed for evaluation of invariance to affine transformations. A further study similar to the one in this paper would be helpful to evaluate those questions. From our experience and the feedback of the study participants, we can however report that users still give good ratings if an object is located at a slightly different position than in the sketch, strong offsets, however, typically lead to poor ratings. Also, slight scale invariance is desirable, rotation invariance however does not seem to be what the average user is looking for.

6.1.1 Possible sources of bias

We note that the selection process of images to be paired with the sketches is potentially biased. A pool of candidate images has been collected using *existing* sketch-based retrieval algorithms and it is hard to assess what images have been ignored in this process.

Potential bias for the study could only be introduced by systematically ignoring images that would be considered good matches to a sketch at least by some human observers. However, the same could be true for *any other automated selection process*. And, also a human-guided selection process would likely be biased: unless we know that humans rate the likeness of sketches and images similarly we would have to involve several subjects in the selection process – whose selection is a sampling problem in itself. More severely, the manual inspection of millions of images will most likely lead to fatigue and frustration, which would bias the selection. Consequently, we believe it is impossible to select a set of images for this task that was arguably unbiased.

We like to explain the consequences of a potentially biased images sample:

1) Most obviously, the selection based on an existing SBIR system could favor this system in the benchmark. We argue that this is very unlikely, since subjects are not aware of how the images have been selected and, indeed, for a few sketches, we noticed negative correlation values between users' rankings and the descriptor's ranking.
2) Our conclusion that humans rank the likeness of sketches and images consistently would have to be limited to the subsample we have actually analyzed, if a significant part of relevant pairs was missing. This is a general problem in any study of this type. However, for the definition of a benchmark this has little consequences, as the conclusion is still true for the sample used.
3) Optimizing descriptor parameters based on the benchmark could overfit to the sample, which is not fully representative of ground truth. This is a common problem in any approach that optimizes an algorithm based on sampled data. We can however report that algorithms that appear improved based on the benchmark indeed behave better in practical applications.

6.1.2 Choice of rating scale

A 5- or 7-point Likert scale is commonly used in psychological experiments – possibly because humans are able to discern just around seven values in many types of experiments [30]. Research on the effect of scales on the outcome of experiments found no statistical difference between 5- and 7-point scales [31]. Using a 7-point scale in our study, none of the subjects reported they disliked the scale or found it more difficult to choose among the pairs that did not match well.

6.2 Benchmark

We propose to use the ranking resulting from averaging scores over all 28 study participants as the benchmark ranking. Our analysis reveals that this ranking can be reasonably used as a benchmark (see Figure 5, right). However, certain characteristics of the

benchmark might need to be adapted depending on the application and the needs of the specific user [8]. Also, there are certainly other valid ways of analyzing the resulting data besides rank correlation. We therefore provide the complete dataset gathered from the user study as an open resource, hoping that this spawns further research on SBIR evaluation.

6.3 Sketching interface

A search result can only be as good as the input, and therefore the performance of current SBIR systems currently largely depends on a user's ability to depict the desired sketch in a "realistic" way. Due to the use of coarse histogram binning, the proposed local feature descriptors are tolerant against offsets in position and orientation of the depicted sketch lines – nevertheless they rely on the fact that the depicted lines exist in the images to be searched. The proposed 2D search from 3D input could be an initial approach to overcome those problems, helping users to generate sketches that are e.g. perspectively correct. Automatic tools for simplification, smoothing, filtering and enforcing symmetry in the sketch could support users in generating "good" sketches. We can also envision approaches where a system "teaches" the user how to draw sketches in order to achieve good retrieval results. Such a system could e.g. filter or mark lines in a sketch that would have little influence on the search results.

6.4 Local features

We found that the local feature size required should be as large as 25% of the image diagonal in order to achieve good query results. Note that with such large window sizes, invariance to translations also suffers – for the large sketches typically drawn by users there is simply not much space left for translating the sketch. This might also explain the good results achieved with global descriptors, which are mostly translation variant. An obvious improvement of the local descriptors – which we left for future work – is making them fully invariant to affine transformations, i.e. rotations (if desired) and scaling. Many possible approaches for this problem have been reported in the image retrieval literature and it remains to be evaluated, whether those approaches can be directly applied to SBIR.

7 Conclusions

Our main conclusion drawn from the results of the user study is that humans consistently agree on the similarity between a simple binary sketch on the one hand and a color image on the other hand. This result has allowed us to define a general benchmark for evaluating the performance of any SBIR system. As shown in this paper, the benchmark cannot only be used to evaluate existing systems but can also be used to optimize new systems by giving a means for measuring the influence of various retrieval system parameters on retrieval performance. We believe that the most important detail concerning our benchmark is that it is grounded on *real users' input*: they need to be satisfied by the query results of a particular system and the score of the proposed benchmark is directly proportional to this property. Interestingly, we found that the rather small difference in benchmark correlation values between the best global and the best local descriptor results in significantly improved experienced retrieval quality (i.e. one can "clearly feel" the difference).

Given the performance of the currently best performing descriptors, we have to admit that those results are still far from optimal – real humans are able to achieve substantially higher correlation values than current systems. We hope that the benchmark provided in this paper will help to stimulate further research on sketch-based image retrieval and make the resulting retrieval systems more comparable.

Acknowledgments

This work was supported in part by a gift from the Apple Research & Technology Support program. We would like to thank Felix Wichmann for insightful discussions about the user study, Ronald Richter and Bert Buchholz for helping implementing the retrieval engine, the reviewers for their constructive feedback and all Flickr users that provided their images under the Creative Commons license.

References

[1] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 1–60, 2008.

[2] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] D. Squire, W. Mueller, H. Mueller, and J. Raki, "Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback," in *Scandinavian Conference on Image Analysis*, 1999, pp. 7–11.

[5] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477.

[6] H. Jégou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," *European Conference on Computer Vision*, pp. 304–317, 2008.

[7] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 25–32.

[8] D. Forsyth, "Benchmarks for storage and retrieval in multimedia databases," *Storage and Retrieval for Media Databases*, pp. 240–247, 2002.

[9] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "An evaluation of descriptors for large-scale image retrieval from sketched feature lines," *Computers & Graphics*, vol. 34, no. 5, pp. 482–498, 2010.

[10] R. Likert, "A technique for the measurement of attitudes." *Archives of Psychology*, vol. 22, no. 140, pp. 1–55, 1932.

[11] T. J. Andrews and D. M. Coppola, "Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments," *Vision Research*, vol. 30, no. 17, pp. 2947–2953, 1999.

[12] F. A. Wichmann, J. Drewes, P. Rosas, and K. R. Gegenfurtner, "Animal detection in natural scenes: Critical features revisited," *Journal of Vision*, vol. 10, no. 4, pp. 1–27, 2010.

[13] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, June 1938.

[14] M. Kendall and J. Gibbons, *Rank correlation methods, 5th ed.* Griffin London, 1990.

[15] N. V. Shirahatti and K. Barnard, "Evaluating image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 955–961.

[16] J. Zobel and A. Moffat, "Inverted files for text search engines," *ACM Computing Surveys*, vol. 38, no. 2, 2006.

[17] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," *European Conference on Computer Vision*, pp. 490–503, 2006.

[18] I. Witten, A. Moffat, and T. Bell, *Managing gigabytes: compressing and indexing documents and images*. Morgan Kaufmann, 1999.

[19] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[21] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, 1986.

[22] A. Chalechale, G. Naghdy, and A. Mertins, "Sketch-based image matching using angular partitioning," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 35, no. 1, pp. 28–41, 2005.

[23] B. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: multimedia content description interface*. John Wiley & Sons Inc, 2002.

[24] P. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[25] D. DeCarlo, A. Finkelstein, S. Rusinkiewicz, and A. Santella, "Suggestive contours for conveying shape," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 848–855, 2003.

[26] T. Judd, F. Durand, and E. H. Adelson, "Apparent ridges for line drawing," *ACM Transactions on Graphics*, vol. 26, no. 3, 2007.

[27] R. Dony, J. Mateer, J. Robinson, and M. Day, "Iconic versus naturalistic motion cues in automated reverse storyboarding," in *Visual Media Production*, 2005, pp. 17–25.

[28] D. B. Goldman, B. Curless, S. M. Seitz, and D. Salesin, "Schematic storyboarding for video visualization and editing," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 862–871, 2006.

[29] J. P. Collomosse, G. McNeill, and Y. Qian, "Storyboard sketches for content based video retrieval," in *IEEE International Conference on Computer Vision*, 2009, pp. 245–252.

[30] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological review*, vol. 63, no. 2, pp. 81–97, 1956.

[31] J. Dawes, "Do data characteristics change according to the number of scale points used?" *International Journal of Market Research*, vol. 50, no. 1, pp. 61–77, 2008.