Data-driven Visual Similarity for Cross-domain Image Matching

Abhinav Shrivastava
Carnegie Mellon University

Tomasz Malisiewicz
MIT

Abhinav Gupta
Carnegie Mellon University

Alexei A. Efros
Carnegie Mellon University

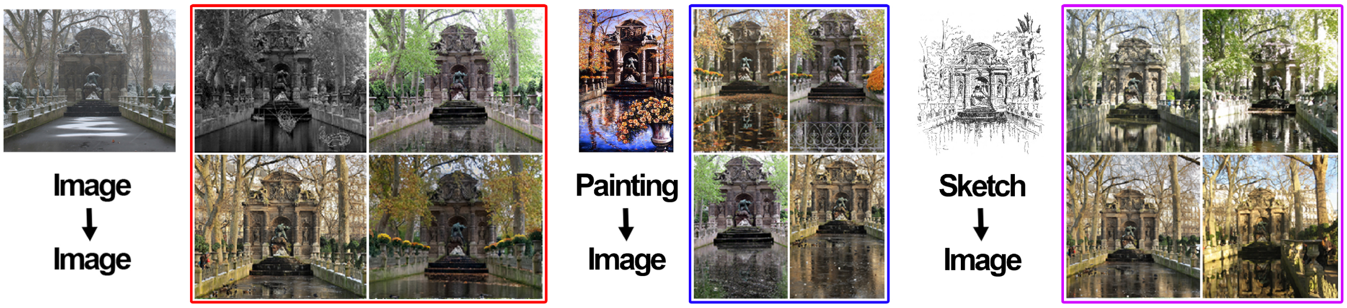


Figure 1: *In this paper, we are interested in defining visual similarity between images across different domains, such as photos taken in different seasons, paintings, sketches, etc. What makes this challenging is that the visual content is only similar on the higher scene level, but quite dissimilar on the pixel level. Here we present an approach that works well across different visual domains.*

Abstract

The goal of this work is to find *visually similar* images even if they appear quite different at the raw pixel level. This task is particularly important for matching images across visual domains, such as photos taken over different seasons or lighting conditions, paintings, hand-drawn sketches, etc. We propose a surprisingly simple method that estimates the relative importance of different features in a query image based on the notion of "data-driven uniqueness". We employ standard tools from discriminative object detection in a novel way, yielding a generic approach that does not depend on a particular image representation or a specific visual domain. Our approach shows good performance on a number of difficult cross-domain visual tasks e.g., matching paintings or sketches to real photographs. The method also allows us to demonstrate novel applications such as *Internet re-photography*, and `painting2gps`. While at present the technique is too computationally intensive to be practical for interactive image retrieval, we hope that some of the ideas will eventually become applicable to that domain as well.

CR Categories: I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—Learning; I.4.10 [Image Processing and Computer Vision]: Image Representation—Statistical;

Keywords: image matching, visual similarity, saliency, image retrieval, paintings, sketches, re-photography, visual memex

Links: DL PDF WEB

1 Introduction

Powered by the availability of Internet-scale image and video collections coupled with greater processing speeds, the last decade has witnessed the rise of data-driven approaches in computer graphics and computational photography. Unlike traditional methods, which employ parametric models to capture visual phenomena, the data-driven approaches use visual data directly, without an explicit intermediate representation. These approaches have shown promising results on a wide range of challenging computer graphics problems, including super-resolution and de-noising [Freeman et al. 2002; Buades et al. 2005; HaCohen et al. 2010], texture and video synthesis [Efros and Freeman 2001; Schodl et al. 2000], image analogies [Hertzmann et al. 2001], automatic colorization [Torralba et al. 2008], scene and video completion [Wexler et al. ; Hays and Efros 2007; Whyte et al. 2009], photo restoration [Dale et al. 2009], assembling photo-realistic virtual spaces [Kaneva et al. 2010; Chen et al. 2009], and even making CG imagery more realistic [Johnson et al. 2010], to give but a few examples.

The central element common to all the above approaches is searching a large dataset to find visually similar matches to a given query – be it an image patch, a full image, or a spatio-temporal block. However, defining a good visual similarity metric to use for matching can often be surprisingly difficult. Granted, in many situations where the data is reasonably homogeneous (e.g., different patches within the same texture image [Efros and Freeman 2001], or different frames within the same video [Schodl et al. 2000]), a simple pixel-wise sum-of-squared-differences (L2) matching works quite well. But what about the cases when the visual content is only similar on the higher scene level, but quite dissimilar on the pixel level? For instance, methods that use scene matching e.g., [Hays and Efros 2007; Dale et al. 2009] often need to match images across different illuminations, different seasons, different cameras, etc. Likewise, retexturing an image in the style of a painting [Hertzmann et al. 2001; Efros and Freeman 2001] requires making visual correspondence between two very different domains – photos and paintings. Cross-domain matching is even more critical for applications such as Sketch2Photo [Chen et al. 2009] and CG2Real [Johnson et al. 2010], which aim to bring domains as different as sketches and CG renderings into correspondence with natural photographs. In all of these cases, pixel-wise matching fares quite poorly, because small perceptual differences can result in arbitrarily large pixel-wise differences. What is needed is a visual metric that can capture the important visual structures that make two images appear similar, yet show robustness to small, unimportant visual details. This is precisely what makes this problem so difficult – the visual similarity algorithm somehow needs to know which visual structures are important for a human observer and which are not.

Currently, the way researchers address this problem is by using var-

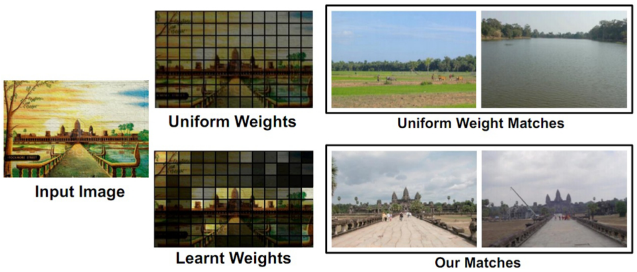


Figure 2: *In determining visual similarity, the central question is which visual structures are important for a human observer and which are not. In the painting above, the brush-strokes in the sky are as thick as those on the ground, yet are perceived as less important. In this paper, we propose a simple, data-driven learning method for determining which parts of a given image are more informative for visual matching.*

ious image feature representations (SIFT [Lowe 2004], GIST [Oliva and Torralba 2006], HoG [Dalal and Triggs 2005], wavelets, etc.) that aim to capture the locally salient (i.e., high gradient and high contrast) parts of the image, while downplaying the rest. Such representations have certainly been very helpful in improving image matching accuracy for a number of applications (e.g., [Hays and Efros 2007; Kaneva et al. 2010; Dale et al. 2009; Johnson et al. 2010]). However, what these features encode are purely local transformations – mapping pixel patches from one feature space into another, independent of the global image content. The problem is that *the same* local feature might be unimportant in one context but crucially important in another. Consider, for example, the painting in Figure 2. In local appearance, the brush-strokes on the alleyway on the ground are virtually the same as the brush-strokes on the sky. Yet, the former are clearly much more informative as to the content of the image than the latter and should be given a higher importance when matching (Figure 2). To do this algorithmically requires not only considering the local features within the context of a given query image, but also having a good way of estimating the importance of each feature with respect to the particular scene's overall visual impression.

What we present in this paper is a very simple, yet surprisingly effective approach to visual matching which is particularly well-suited for matching images across different domains. We do not propose any new image descriptors or feature representations. Instead, given an image represented by some features (we will be using the spatially-rigid HoG [Dalal and Triggs 2005] descriptor for most of this paper), the aim is to focus the matching on the features that are the most visually important *for this particular image*. The central idea is the notion of "data-driven uniqueness". We hypothesize, following [Boiman and Irani 2007], that the important parts of the image are those that are more unique or rare within the visual world (represented here by a large dataset). For example, in Figure 2, the towers of the temple are very unique, whereas the wispy clouds in the sky are quite common. However, since the same local features could represent very different visual content depending of context, unlike [Boiman and Irani 2007], our notion of uniqueness is *scene-dependent* i.e., each query image decides what is the best way to weight its constituent parts. Figure 2 demonstrates the difference between image matching using a standard uniform feature weighting vs. our uniqueness-based weighting.

We operationalize this data-driven uniqueness by using ideas from machine learning – training a discriminative classifier to discover which parts of an image are most discriminative *in relationship to the rest of the dataset*. This simple approach results in visual matching that is surprisingly versatile and robust. By focusing on the globally salient parts of the image, the approach can be successfully used for generic cross-domain matching without making any domain-specific changes, as shown on Figure 1. The rest of the paper is organized as follows: we first give a brief overview of the related work (Section 1.1), then describe our approach in detail (Section 2), present an evaluation on several public datasets (Section 3), and finally show some of the applications that our algorithm makes possible (Section 4).

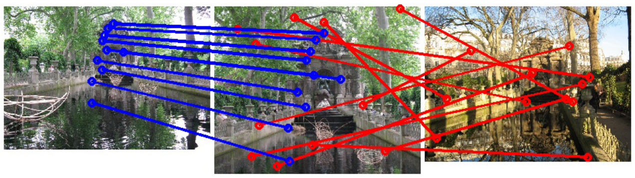


Figure 3: *Example of image matching using the SIFT descriptor. While SIFT works very well at matching fine image structure (left), it fails miserably when there is too much local change, such as a change of season (right).*

1.1 Background

In general, visual matching approaches can be divided into three broad classes, with different techniques tailored for each:

Exact matching: For finding more images of the exact same physical object (e.g., a Pepsi can) or scene (e.g., another photo of Eiffel Tower under similar illumination), researchers typically use the general bag-of-words paradigm introduced by the Video Google work [Sivic and Zisserman 2003], where a large histogram of quantized local image patches (usually encoded with the SIFT descriptor [Lowe 2004]) is used for image retrieval. This paradigm generally works extremely well (especially for heavily-textured objects), and has led to many successful applications such as Google Goggles. However, these methods usually fail when tasked with finding *similar*, but not identical objects (e.g., try using Google Goggles *app* to find a cup, or a chair). This is because SIFT, being a local descriptor, captures the minute details of a particular object well, but not its overall global properties (as seen in Figure 3).

Approximate matching: The task of finding images that are merely "visually similar" to a query image is significantly more difficult and none of the current approaches can claim to be particularly successful. Most focus on employing various image representations that aim to capture the important, salient parts of the image. Some of the popular ones include the GIST [Oliva and Torralba 2006] descriptor, the Histogram of Gradients (HoG) descriptor [Dalal and Triggs 2005], various other wavelet- and gradient-based decompositions, or agglomerations, such as the spatial pyramid [Lazebnik et al. 2009] of visual words. Also related is the vast field of Content-Based Image Retrieval (CBIR) (see [Datta et al. 2008] for overview). However, in CBIR the goals are somewhat different: the aim is to retrieve *semantically-relevant* images, even if they do not appear to be visually similar (e.g., a steam-engine would be considered semantically very similar to a bullet train even though visually there is little in common). As a result, most modern CBIR methods combine visual information with textual annotations and user input.

Cross-domain matching: A number of methods exists for matching between particular domains, such as sketches to photographs (e.g., [Chen et al. 2009; Eitz et al. 2010]), drawings/paintings to photographs (e.g., [Russell et al. 2011]), or photos under different illuminants (e.g., [Chong et al. 2008]), etc. However these typically present very domain-specific solutions that do not easily generalize across multiple domains. Of the general solutions, the most ambitious is work by Shechtman and Irani [2007], which proposes to describe an image in terms of local self-similarity descriptors that are invariant across visual domains. This work is complementary to ours since it focuses on the design of a cross-domain local descriptor, while we consider relative weighting between the descriptors for a given image, so it might be interesting to combine both.

Within the text retrieval community, the *tf-idf* normalization [Baeza-Yates and Ribeiro-Neto 1999] used in the bag-of-words approaches shares the same goals as our work – trying to re-weight the different features (words in text, or "visual words" in images [Sivic and Zisserman 2003]) based on their relative frequency. The main difference is that in *tf-idf*, each word is re-weighted independently of all the others, whereas our method takes the interactions between all of the features into account.

Most closely related to ours are approaches that try to learn the statistical structure of natural images by using large unlabeled image sets, as a way to define a better visual similarity. In the context of image retrieval, Hoiem et al. [2004] estimate the unconditional probability density of images off-line and use it in a Bayesian framework to find close matches; Tieu and Viola [2004] use boosting at query-time to discriminatively learn query-specific features. However, these systems require multiple positive query images and/or user guidance, whereas most visual matching tasks that we are interested in need to work automatically and with only a single input image. Fortunately, recent work in visual recognition has shown that it's possible to train a discriminative classifier using a *single positive instance* and a large body of negatives [Wolf et al. 2009; Malisiewicz et al. 2011], provided that the negatives do not contain any images similar to the positive instance. In this work, we adapt this idea to image retrieval, where one cannot guarantee that the "negative set" will not contain images similar to the query (on the contrary, it most probably will!). What we show is that, surprisingly, this assumption can be relaxed without adversely impacting the performance.

2 Approach

The problem considered in this paper is the following: how to compute visual similarity between images which would be more consistent with human expectations. One way to attack this is by designing a new, more powerful image representation. However, we believe that existing representations are already sufficiently powerful, but that the main difficulty is in developing the right similarity distance function, which can "pick" which parts of the representation are most important for matching. In our view, there are two requirements for a good visual similarity function: 1) It has to focus on the content of the image (the "what"), rather that the style (the "how") e.g., the images on Figure 1 should exhibit high visual similarity despite large pixel-wise differences. 2) It should be *scene-dependent*, that is, each image should have its own unique similarity function that depends on its global content. This is important since the same local feature can represent vastly different visual content, depending on what else is depicted in the image.

2.1 Data-driven Uniqueness

The visual similarity function that we propose is based on the idea of "data-driven uniqueness". We hypothesize that what humans find important or salient about an image is somehow related to how unusual or unique it is. If we could re-weight the different elements of an image based on how unique they are, the resulting similarity function would, we argue, answer the requirements of the previous section. However, estimating "uniqueness" of a visual signal is not at all an easy task, since it requires a very detailed model of our entire visual world, since only then we can know if something is truly unique. Therefore, instead we propose to compute uniqueness in a data-driven way — against a very large dataset of randomly selected images.

The basic idea behind our approach is that the features of an image that exhibit high "uniqueness" will also be the features that would best discriminate this image (the positive sample) against the rest of the data (the negative samples). That is, we are able to map the highly complex question of visual similarity into a fairly standard problem in discriminative learning. Given some suitable way of representing an image as a vector of features, the result of the discriminative learning is a set of weights on these features that provide for the best discrimination. We can then use these same weights to compute visual similarity. Given the learned, query-dependent weight vector $\mathbf{w}_q$, the visual similarity between a query image I_q and any other image/sub-image I_i can be defined simply as:

$$S(I_q, I_i) = {\mathbf{w}_q}^T \mathbf{x}_i \quad (1)$$

where $\mathbf{x}_i$ is I_i's extracted feature vector.

To learn the feature weight vector which best discriminates an image from a large "background" dataset, we employ the linear Support Vector Machine (SVM) framework. We set up the learning problem following [Malisiewicz et al. 2011] which has demonstrated that a linear SVM can generalize even with a single positive example, provided that a very large amount of negative data is available to "constrain the solution". However, whereas in [Malisiewicz et al. 2011] the negatives are guaranteed not to be members of the positive class (that is why they are called negatives), here this is not the case. The "negatives" are just a dataset of images randomly sampled from a large Flickr collection, and there is no guarantee that some of them might not be very similar to the "positive" query image. Interestingly, in practice, this does not seem to hurt the SVM, suggesting that this is yet another new application where the SVM formalism can be successfully applied.

The procedure described above should work with any sufficiently powerful image feature representation. For the majority of our experiments in this paper, we have picked the Histogram of Oriented Gradients (HOG) template descriptor [Dalal and Triggs 2005], due to its good performance for a variety of tasks, its speed, robustness, adaptability to sliding window search, and popularity in the community. We also show how our learning framework can be used with Dense-SIFT (D-SIFT) template descriptor in Section 2.4.

To visualize how the SVM captures the notion of data-driven uniqueness, we performed a series of experiments with simple, synthetic data. In the first experiment, we use simple synthetic figures (a combination of circles and rectangles) as visual structures on the query image side. Our negative world consists of just rectangles of multiple sizes and aspect ratios. If everything works right, using the SVM-learned weights should downplay the features (gradients in HoG representation) generated from the rectangle and increase the weights of features generated by the circle, since they are more unique. We use the HoG visualization introduced by [Dalal and Triggs 2005] which displays the learned weight vector as a gradient distribution image. As Figure 4(a) shows, our approach indeed suppresses the gradients generated by the rectangle.

One of the key requirements of our approach is that it should be able to extract visually important regions even when the images are from different visual domains. We consider this case in our next experiment, shown on Figure 4(b). Here the set of negatives includes two domains – black-on-white rectangles and white-on-black rectangles. By having the negative set include both domains, our approach should downplay any domain-dependent idiosyncrasies both from the point of view of the query and target domains. Indeed, as Figure 4(b) shows, our approach is again able to extract the unique structures corresponding to circles while downplaying the gradients generated due to rectangles, in a domain-independent way.

We can also observe this effect on real images. The Venice bridge painting shown in Figure 5 initially has high gradients for building boundaries, the bridge and the boats. However, since similar building boundaries are quite common, they occur a lot in the randomly sampled negative images and hence, their weights are reduced.

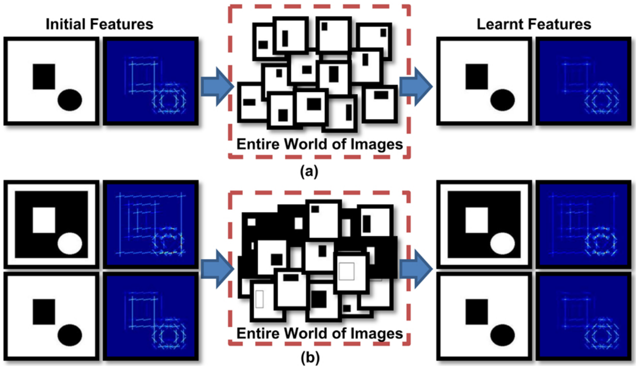


Figure 4: *Synthetic example of learning data-driven "uniqueness". In each case, our learned similarity measure boosts the gradients belonging to the circle because they are more unique with respect to a synthetic world of rectangle images.*

2.2 Algorithm Description

We set up the learning problem using a single positive and a very large negative set of samples similar to [Malisiewicz et al. 2011]. Each query image (I_q) is represented with a rigid grid-like HoG feature template ($\mathbf{x}_q$). We perform binning with sizing heuristics which attempt to limit the dimensionality of ($\mathbf{x}_q$) to roughly $4{-}5K$, which amounts to $\sim$150 cells for HoG template. To add robustness to small errors due to image misalignment, we create a set of extra positive data-points, $\mathcal{P}$, by applying small transformations (shift, scale and aspect ratio) to the query image I_q, and generating $\mathbf{x}_i$ for each sample. Therefore, the SVM classifier is learned using I_q and $\mathcal{P}$ as positive samples, and a set containing millions of sub-images $\mathcal{N}$ (extracted from $10,000$ randomly selected Flickr images), as negatives. Learning the weight vector $\mathbf{w}_q$ amounts to minimizing the following convex objective function:

$$L(\mathbf{w}_q) = \sum_{\mathbf{x}_i \in \mathcal{P} \cup I_q} h(\mathbf{w}_q^T \mathbf{x}_i) + \sum_{\mathbf{x}_j \in \mathcal{N}} h(-\mathbf{w}_q^T \mathbf{x}_j) + \lambda ||\mathbf{w}_q||^2 \quad (2)$$

We use `LIBSVM` [Chang and Lin 2011] for learning $\mathbf{w}_q$ with a common regularization parameter $\lambda = 100$ and the standard hinge loss function $h(x) = \max(0, 1 - x)$. The hinge-loss allows us to use the hard-negative mining approach [Dalal and Triggs 2005] to cope with millions of negative windows because the solution only depends on a small set of negative support vectors. In hard-negative mining, one first trains an initial classifier using a small set of training examples, and then uses the trained classifier to search the full training set exhaustively for false positives ('hard examples'). Once sufficient number of hard negatives are found in the training set, one retrains the classifier $\mathbf{w}_q$ using this set of hard examples. We alternate between learning $\mathbf{w}_q$ given a current set of hard-negative examples, and mining additional negative examples using the current $\mathbf{w}_q$ as in [Dalal and Triggs 2005]. For all experiments in this paper, we use 10 iterations of hard-mining procedure; with each iteration requiring more time than the previous one because it becomes harder to find hard-negatives as the classifier improves. Empirically, we found that more than 10 iterations did not provide enough improvement to justify the run-time cost.

The standard sliding window setup [Dalal and Triggs 2005] is used to evaluate all the sub-windows of each image. For this, the trained classifier is convolved with the HoG feature pyramid at multiple scales for each image in the database. The number of pyramid levels controls the size of possible detected windows in the image. We use simple non-maxima suppression to remove highly-overlapping redundant matches. While the use of sub-window search is expensive, we argue that it is crucial to good image matching for the following reasons. First, it allows us to see millions of negative examples during training from a relatively small number of images $(10,000)$. But more importantly, as argued by [Hoiem et al. 2004], sub-window search is likely to dramatically increase the number of good matches over the traditional full-image retrieval techniques.

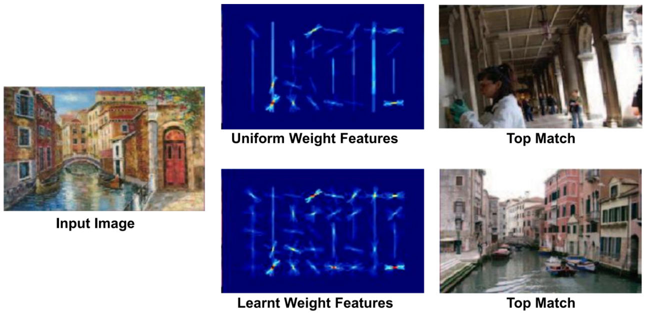


Figure 5: *Learning data-driven uniqueness: Our approach down-weighs the gradients on the buildings since they are not as rare as the circular gradients from the bridge.*

2.3 Relationship to Saliency

We found that our notion of data-driven uniqueness works surprisingly well as a proxy for predicting image saliency ("where people look") – a topic of considerable interest to computer graphics. We ran our algorithm on the human gaze dataset from Judd et al. [2009], using a naive mapping from learned HoG weights to predicted pixel saliency by spatially summing these weights followed by normalization. Figure 6 compares our saliency prediction against standard saliency methods (summarized in [Judd et al. 2009]). While our score of 74% (mean area under ROC curve) is below [Judd et al. 2009] who are the top performers at 78% (without center prior), we beat most classic saliency methods such as Itti et al. [2000] who only obtained 62%. After incorporating a simple gaussian center prior, our score raises to 81.9%, which is very close to 83.8% of [Judd et al. 2009].

2.4 Other Features

Our framework should be able to work with any rigid grid-like image representation where the template captures feature distribution in some form of histogram of high-enough dimensionality. We performed preliminary experiments using the dense SIFT (D-SIFT) template descriptor (similar to [Lazebnik et al. 2009]) within our framework for the task of Sketch-to-Image Matching (Section 3.2). The query sketch (I_q) was represented with a feature template ($\mathbf{x}_q$) of D-SIFT and sizing heuristics (Section 2.2) produced $\sim$35 cells for the template (128 dimensions per cell). Figure 10 demonstrates the results of these preliminary experiments, where our learning framework improves the performance of D-SIFT baseline (without learning) indicating that our algorithm can be adapted to a different feature representation.

3 Experimental Validation

To demonstrate our approach, we performed a number of image matching experiments on different image datasets, comparing against the following popular baseline methods:

Tiny Images: Following [Torralba et al. 2008], we re-size all images to 32x32, stack them into 3072-D vectors, and compare them using Euclidean distance.

GIST: We represent images with the GIST [Oliva and Torralba 2006] descriptor, and compare them with the Euclidean distance.

BoW: We compute a Bag-of-Words representation for each image using vector-quantized SIFT descriptors [Lowe 2004] and compare the visual word histograms (with *tf-idf* normalization) as in [Sivic and Zisserman 2003].

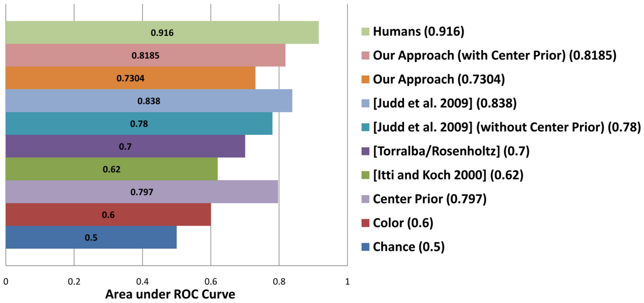


Figure 6: *The concept of data-driven uniqueness can also be used as a proxy to predict saliency for an image. Our approach performs better than individual features (such as Itti et al. and Torralba/Rosenholtz, see [Judd et al. 2009]) and comparable to Judd et al. [2009].*

Spatial Pyramid: For each image, we compute spatial pyramid [Lazebnik et al. 2009] representation with 3 pyramid levels using Dense-SIFT descriptors of 16x16 pixel patches computed over a grid with spacing of 8 pixels. We used vocabulary of 200 visual words. The descriptors are compared using histogram intersection pyramid matching kernels as described in [Lazebnik et al. 2009].

Normalized-HoG (N-HoG): We represent each image using the same HoG descriptor as our approach, but instead of learning a query-specific weight vector, we match images directly in a nearest-neighbor fashion. We experimented with different similarity metrics and found a simple normalized HoG (N-HoG) to give the best performance. The N-HoG weight vector is defined as a zero-centered version of the query's HoG features $\mathbf{w}_q = \mathbf{x}_q - mean(\mathbf{x}_q)$. Matching is performed using Equation 1, by replacing the learned weight vector with N-HoG weight vector.

In addition, we also compare our algorithm to Google's recently released Search-by-Image feature. It should be noted that the retrieval dataset used by Google is orders of magnitude larger than the tens of thousands of images typically used in our datasets, so this comparison is not quite fair. But while Google's algorithm shows a reasonable performance in retrieving landmark images with similar illumination, season and viewpoint, it does not seem to adapt well to photos taken under different lighting conditions or photos from different visual domains such as sketches and paintings (see Figure 9).

3.1 Image-to-Image Matching

While image retrieval is not the goal of this paper, the CBIR community has produced a lot of good datasets that we can use for evaluation. Here we consider the instance retrieval setting using the *INRIA Holidays* dataset introduced by Jégou et al. [2008] and one million random distractor Flickr images from [Hays and Efros 2007] to evaluate performance. The goal is to measure the quality of the top matching images when the exact instances are present in the retrieval dataset. For evaluation, we follow [Jégou et al. 2008] and measure the quality of rankings as the true positive rate from the list of top $k = 100$ matches as a function of increasing dataset size. Since the average number of true positives is very small for the Holidays dataset, we also perform the evaluation with smaller k. We compare our approach against GIST, Tiny Images and Spatial Pyramid baselines described in Section 3 on 50 random Holidays query images and evaluate the top 5 and 100 matches for the same dataset sizes used in [Jégou et al. 2008].

Table 1 demonstrates the robustness of our algorithm to adding distractor images – the true positives rate only drops from 69% to 62% when we add 1M distractors (which is of similar order as in [Jégou et al. 2008]), outperforming the state-of-art spatial pyramid matching [Lazebnik et al. 2009]. It is important to note that even after drastically reducing the ranks under consideration from the top 100 to just the top 5, our rate of true positives drops by only 3% (which attests to the quality of our rankings). For a dataset of one million images and a short-list of 100, [Jégou et al. 2008] return 62% true positives which is only slightly better than our results; however, their algorithm is designed for instance recognition, whereas our approach is applicable to a broad range of cross-domain visual tasks.

Top-5				
Dataset Size	1,490	11,490	101,490	1,001,490
GIST	0.0106	0.0106	0.0106	0.0106
Tiny Images	0.0106	0.0106	0.0106	0.0106
Spatial Pyramid	0.3417	0.3063	0.2471	0.1967
Our Approach	**0.6588**	**0.6393**	**0.5890**	**0.5836**
Top-100				
Dataset Size	1,490	11,490	101,490	1,001,490
GIST	0.1921	0.1417	0.1417	0.1417
Tiny Images	0.0713	0.0518	0.0518	0.0518
Spatial Pyramid	0.4888	0.415	0.3448	0.2792
Our Approach	**0.6874**	**0.6874**	**0.6619**	**0.6150**

Table 1: *Instance retrieval in Holidays dataset + Flickr1M. We report the mean true positive rate from the top-k image matches as a function of increasing dataset size (averaged across a set of 50 Holidays query images).*

3.2 Sketch-to-Image Matching

Matching sketches to images is a difficult cross-domain visual similarity task. While most current approaches use specialized methods tailored to sketches, here we apply exactly the same procedure as before, without any changes. We collected a dataset of 50 sketches (25 cars and 25 bicycles) to be used as queries (our dataset includes both amateur sketches from the internet as well as freehand sketches collected from non-expert users). The sketches were used to query into the PASCAL VOC dataset [Everingham et al. 2007], which is handy for evaluation since all the car and bicycle instances have been labeled. Figure 8(top) show some example queries and the corresponding top retrieval results for our approach and the baselines. It can be seen that our approach not only outperforms all of the baselines, but returns images showing the target object in a very similar pose and viewpoint as the query sketch.

For quantitative evaluation, we compared how many car and bicycle images were retrieved in the top-K images for car and bicycle sketches respectively. We used the bounded mean Average Precision (mAP) metric used by [Jégou et al. 2008] [1]. We evaluated the performance of our approach (using HoG and D-SIFT) as a function of dataset size and compare it with the multiple baselines, showing the robustness of our approach to the presence of distractors. For each query, we start with all images of the target class from the dataset, increase the dataset size by adding 1000, 5000 images and finally the entire PASCAL VOC 2007 dataset. Figure 10(a) and (b) show mAP as a function of dataset size for cars and bicycles, respectively. For the top 150 matches, we achieve a mAP of 67% for cars and 54% for bicycles (for Learnt-HoG). We also ran our algorithm on the Sketch-Based Image Retrieval (SBIR) Benchmark Dataset [Eitz et al. 2010]. For the top 20 similar images ranked by users, we retrieve 51% of images as top 20 matches, compared 63% using a sketch-specific method of [Eitz et al. 2010]

3.3 Painting-to-Image Matching

As another cross-domain image matching evaluation, we measured the performance of our system on matching paintings to images. Retrieving images similar to paintings is an extremely difficult

[1] Maximum recall is bounded by the number of images being retrieved. For example, if we consider only top-150 matches the maximum true positives would be 150 images

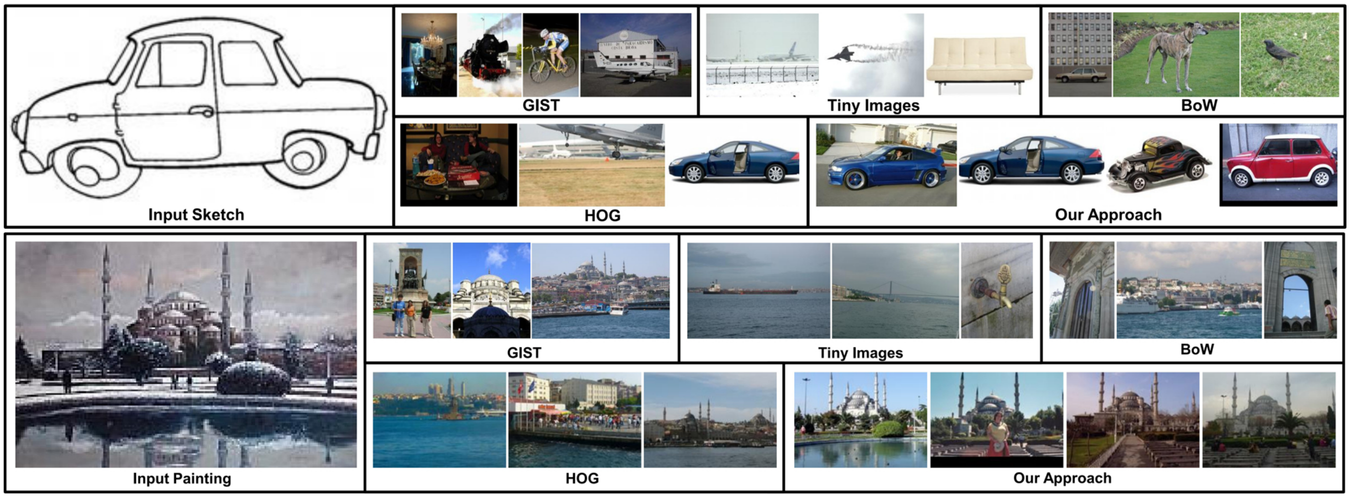


Figure 7: *Qualitative comparison of our approach against baselines for Sketch-to-Image and Painting-to-Image matching.*

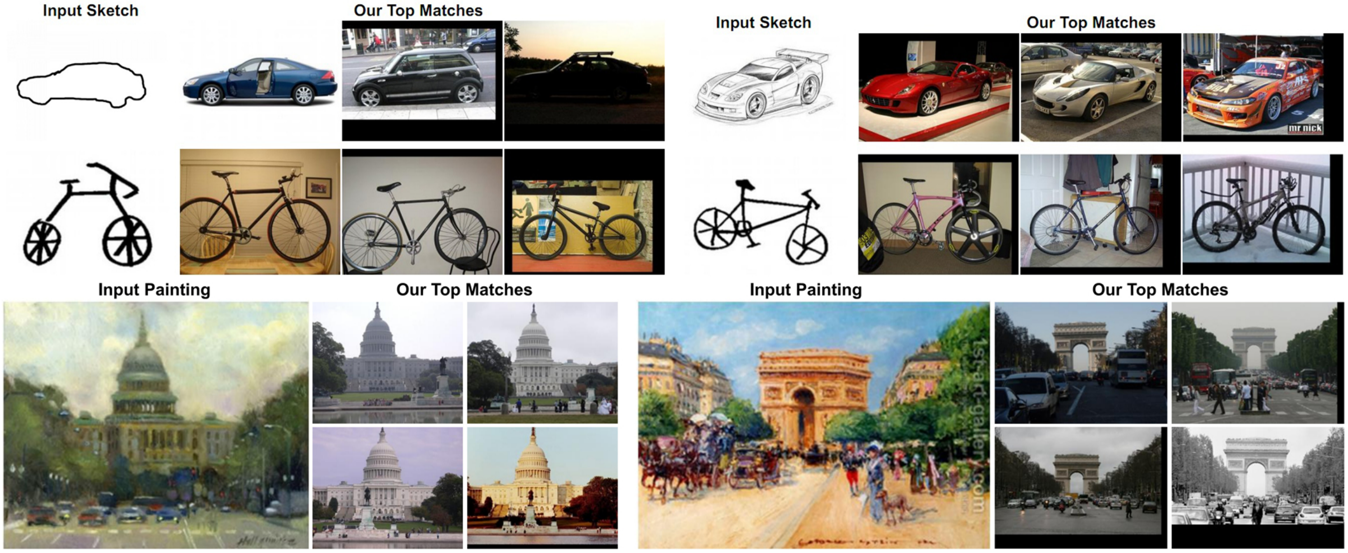


Figure 8: *A few more qualitative examples of top-matches for sketch and painting queries.*

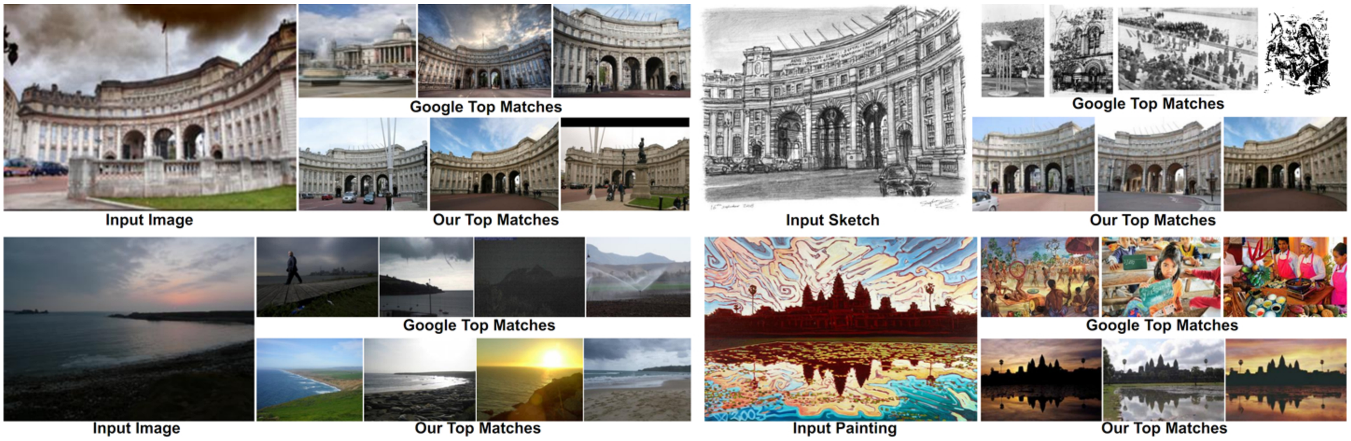
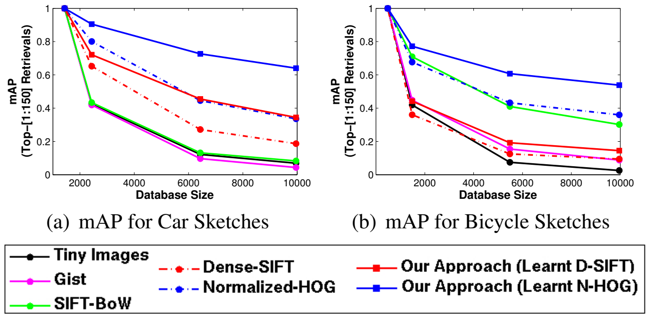


Figure 9: *Qualitative comparison of our approach with Google's 'Search–by–Image' feature. While our approach is robust to illumination changes and performs well across different visual domains, Google image search fails completely when the exact matches are not in the database.*



(a) mAP for Car Sketches (b) mAP for Bicycle Sketches

Figure 10: *Sketch-to-Image evaluation. We match car/bicycle sketches to sub-images in the PASCAL VOC 2007 dataset and measure performance as the number of distractors increases.*

problem because of the presence of strong local gradients due to brush strokes (even in the regions such as sky). For this experiment, we collected a dataset of 50 paintings of outdoor scenes in a diverse set of painting styles geographical locations. The retrieval set was sub-sampled from the 6.4M GPS-tagged Flickr images of [Hays and Efros 2008]. For each query, we created a set of 5,000 images randomly sampled within a 50 mile radius of each painting's location (to make sure to catch the most meaningful distractors), and 5,000 random images. Qualitative examples can be seen in Figure 7.

4 Applications

Our data-driven visual similarity measure can be used to improve many existing matching-based application, as well as facilitate new ones. We briefly discuss a few of them here.

4.1 Better Scene Matching for Scene Completion

Data-driven Scene Completion has been introduced by [Hays and Efros 2007]. However, their scene matching approach (using the GIST descriptor) is not always able to find the best matches automatically. Their solution is to present the user with the top 20 matches and let him find the best one to be used for completion. Here we propose to use our approach to automate scene completion, removing the user from the loop. To evaluate the approach, we used the 78 query images from the scene completion test set [Hays and Efros 2007] along with the top 160 results retrieved by them. We use our algorithm to re-rank these 160 images and evaluate both the quality of scene matches and scene completions against [Hays and Efros 2007].

Figure 11 shows a qualitative result for the top match using our approach as compared to the top match from the GIST+ features used by [Hays and Efros 2007]. To compute quantitative results, we performed two small user studies. In the first study, for each query image participants were presented with the best scene match using our approach, [Hays and Efros 2007] and tiny-images [Torralba et al. 2008]. Participants were asked to select the closest scene match out of the three options. In the second study, participants were presented with automatically completed scenes using the top matches for all three algorithms, and were asked to select the most convincing/compelling completion. The order of presentation of queries as well as the order of the three options were randomized. Overall, for the first task of scene matching, the participants preferred our approach in 51.4% cases as opposed 27.6% for [Hays and Efros 2007] and 21% for Tiny-Images. For the task of automatic scene completion, our approach was found to be more convincing in 47.3% cases as compared to 27.5% for [Hays and Efros 2007] and 25.2% for Tiny-Images. The standard-deviation of user responses for most of the queries were surprisingly low.

4.2 Internet Re-photography

We were inspired by the recent work on computational rephotography [Bae et al. 2010], which allows photographers to take modern photos that match a given historical photograph. However, the approach is quite time-consuming, requiring the photographer to go "on location" to rephotograph a particular scene. What if, instead of rephotographing ourselves, we could simply find the right modern photograph online? This seemed like a perfect case for our cross-domain visual matching, since old and new photographs look quite different and would not be matched well by existing approaches.

We again use the 6.4M geo-tagged Flickr images of [Hays and Efros 2007], and given an old photograph as a query, we use our method to find its top matches from a pre-filtered set of 5,000 images closest to the old photograph's location (usually at least the city or region is known). Once we have an ordered set of image matches, the user can choose one of the top five matches to generate the best old/new collage. Re-photography examples can be seen in Figure 12.

4.3 Painting2GPS

Wouldn't it be useful if one could automatically determine from which location a particular painting was painted? Matching paintings to real photos from a large GPS-tagged collection allows us to estimate the GPS coordinates of the input painting, similar to the approach of [Hays and Efros 2008]. We call this application `painting2GPS`. We use painting-to-image matching as described in Section 3.3, and then find the GPS distribution using the algorithm in [Hays and Efros 2008]. Qualitative painting2GPS examples overlayed onto Google-map can be seen in Figure 13.

4.4 Visual Scene Exploration

Having a robust visual similarity opens the door to interesting ways of exploring and reasoning about large visual data. In particular, one can construct a *visual memex* graph (using the terminology from [Malisiewicz and Efros 2009]), whose nodes are images/subimages, and edges are various types of associations, such as visual similarity, context, etc. By visually browsing this memex graph, one can explore the dataset in a way that makes explicit the ways in which the data is interconnected. Such graph browsing visualizations have been proposed for several types of visual data, such as photos of a 3D scene [Snavely et al. 2008], large collections of outdoor scenes [Kaneva et al. 2010], and faces [Kemelmacher-Shlizerman et al. 2011]. Here we show how our visual similarity can be used to align photos of a scene and construct a movie. Given a set of 200 images automatically downloaded from Flickr using keyword search (e.g., "Medici Fountain Paris"), we compute an all-to-all matrix of visual similarities that represents our visual memex graph. Note that because we are using scanning window matching on the detection side, a zoomed-in scene detail can still match to a wide-angle shot as seen on Figure 14 (top). Other side-information can also be added to the graph, such as the relative zoom factor, or similarity in season and illumination (computed from photo timestamps). One can now interactively browse through the graph, or create a visual memex movie showing a particular path from the data, as shown on Figure 14 (bottom), and in supplementary video.

5 Limitations and Future Work

The two main failure modes of our approach are illustrated on Figure 15. In the first example (left), we fail to find a good match due to the relatively small size of our dataset (10,000 images) compared to Google's billions of indexed images. In the second example (right), the query scene is so cluttered that it is difficult for any algorithm to decide which parts of the scene – the car, the people on sidewalk, the building in the background – it should focus on. Addressing this

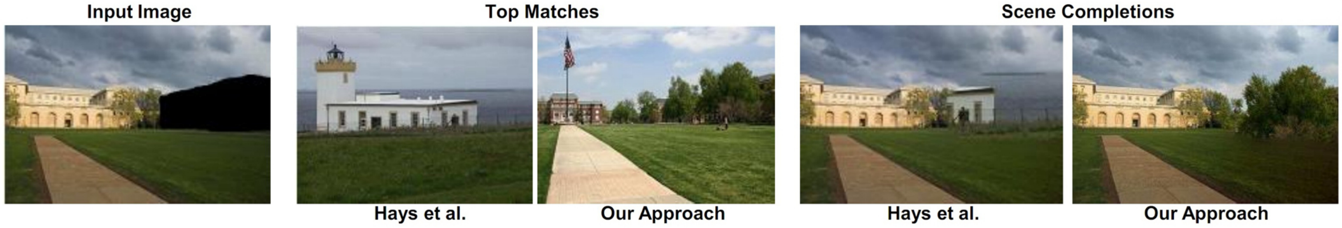


Figure 11: *Qualitative examples of scene completion using our approach and [Hays and Efros 2007].*

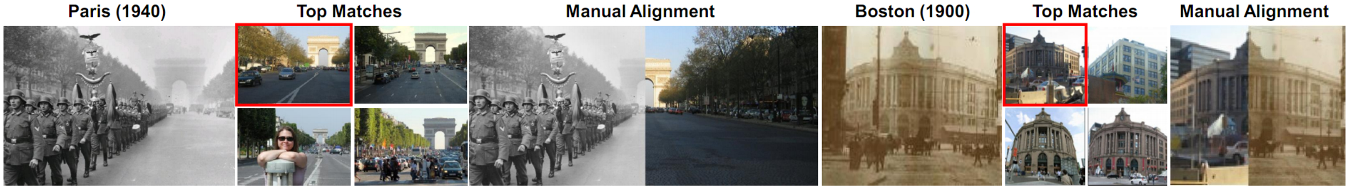


Figure 12: *Internet Re-photography. Given an old photograph, we harness the power of large Internet datasets to find visually similar images. For each query we show the top* 4 *matches, and manually select one of the top matches and create a manual image alignment.*

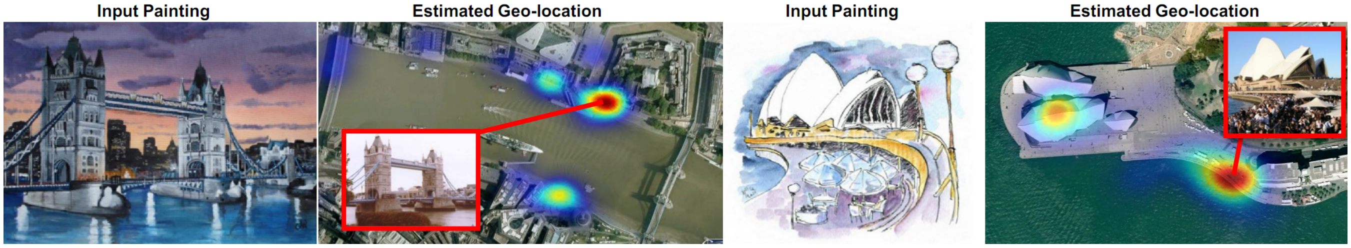


Figure 13: *Painting2GPS Qualitative Examples. In these two painting examples (Tower Bridge in London and the Sydney Opera House), we display estimated GPS location of the painting as a density map overlaid onto Google-map, and the top matching image.*

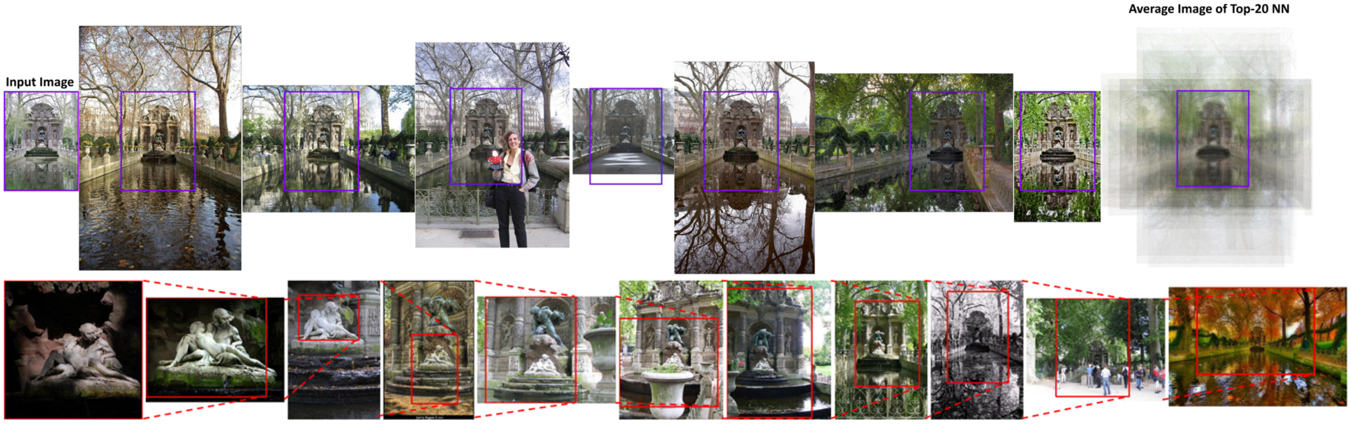


Figure 14: *Visual Scene Exploration. (Top): Given an input image, we show the top matches, aligned by the retrieved sub-window. The last image shows the average of top 20 matches. (Bottom): A visualization of the memex-graph tour through the photos of the Medici Fountain.*

issue will likely require deeper level of image understanding than is currently available.

Speed remains the central limitation of the proposed approach, since it requires training an SVM (with hard-negative mining) at query time. While we developed a fast, parallelized implementation that takes under three minutes per query on a 200-node cluster, this is still too slow for many practical applications at this time. We are currently investigating ways of sharing the computation by precomputing some form of representation for the space of query images ahead of time. However, even in its present form, we believe that the increased computational cost of our method is a small price to pay for the drastic improvements in quality of visual matching.

Acknowledgments: The authors would like to thank Martial Hebert and Bryan Russell for many helpful discussions. This work was supported by ONR Grant N000141010766, MSR-CMU Center for Computational Thinking and Google. Image credits: Charalampos Laskaris, Carol Williams, Claudio Conforti, Eddie Wong, Edson Campos, Prof. Hall Groat II, Kathleen Brodeur, Moira Munro, Matt Wyatt, Keith Hornblower, Don Amadio (Scrambled Eggz Productions), The Stephen Wilthsire Gallery, `www.daydaypaint.com`, The Art Renewal Center and Bundesarchiv.

References

BAE, S., AGARWALA, A., AND DURAND, F. 2010. Computational rephotography. *ACM Trans. Graph. 29* (July), 24:1–24:15.

BAEZA-YATES, R. A., AND RIBEIRO-NETO, B. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing.

BOIMAN, O., AND IRANI, M. 2007. Detecting irregularities in images and in video. In *IJCV*.

BUADES, A., COLL, B., AND MOREL, J.-M. 2005. A non-local algorithm for image denoising. In *CVPR*.

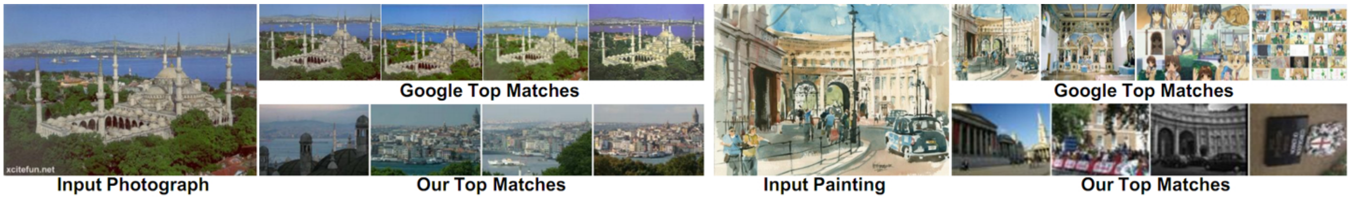


Figure 15: *Typical failure cases. (Left): relatively small dataset size, compared to Google. (Right): too much clutter in the query image.*

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*.

Chen, T., Cheng, M.-M., Tan, P., Shamir, A., and Hu, S.-M. 2009. Sketch2photo: internet image montage. *ACM Trans. Graph. 28*.

Chong, H., Gortler, S., and Zickler, T. 2008. A perception-based color space for illumination-invariant image processing. In *Proceedings of SIGGRAPH*.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*.

Dale, K., Johnson, M. K., Sunkavalli, K., Matusik, W., and Pfister, H. 2009. Image restoration using online photo collections. In *ICCV*.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*.

Efros, A. A., and Freeman, W. T. 2001. Image quilting for texture synthesis and transfer. In *SIGGRAPH*, Computer Graphics Proceedings, Annual Conference Series.

Eitz, M., Hildebrand, K., Boubekeur, T., and Alexa, M. 2010. Sketch-based image retrieval: benchmark and bag-of-features descriptors. *IEEE TVCG*.

Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J., and Zisserman, A., 2007. The PASCAL Visual Object Classes Challenge.

Freeman, W. T., Jones, T. R., and Pasztor, E. C. 2002. Example-based super-resolution. *IEEE Computer Graphics Applications*.

HaCohen, Y., Fattal, R., and Lischinski, D. 2010. Image upsampling via texture hallucination. In *ICCP*.

Hays, J., and Efros, A. A. 2007. Scene completion using millions of photographs. *ACM Transactions on Graphics (SIGGRAPH)*.

Hays, J., and Efros, A. A. 2008. im2gps: estimating geographic information from a single image. In *CVPR*.

Hertzmann, A., Jacobs, C., Oliver, N., Curless, B., and Salesin, D. 2001. Image analogies. In *SIGGRAPH*.

Hoiem, D., Sukthankar, R., Schneiderman, H., and Huston, L. 2004. Object-based image retrieval using the statistical structure of images. In *CVPR*.

Itti, L., and Koch, C. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*.

Jégou, H., Douze, M., and Schmid, C. 2008. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*.

Johnson, M. K., Dale, K., Avidan, S., Pfister, H., Freeman, W. T., and Matusik, W. 2010. CG2real: Improving the realism of computer generated images using a large collection of photographs. *IEEE TVCG*.

Judd, T., Ehinger, K., Durand, F., and Torralba, A. 2009. Learning to predict where humans look. In *ICCV*.

Kaneva, B., Sivic, J., Torralba, A., Avidan, S., and Freeman, W. T. 2010. Infinite images: Creating and exploring a large photorealistic virtual space. *Proceedings of the IEEE*.

Kemelmacher-Shlizerman, I., Shechtman, E., Garg, R., and Seitz, S. M. 2011. Exploring photobios. In *SIGGRAPH*.

Lazebnik, S., Schmid, C., and Ponce, J. 2009. Spatial pyramid matching. In *Object Categorization: Computer and Human Vision Perspectives*. Cambridge University Press.

Lowe, D. 2004. Distinctive image features from scale-invariant keypoints. *IJCV*.

Malisiewicz, T., and Efros, A. A. 2009. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS*.

Malisiewicz, T., Gupta, A., and Efros, A. A. 2011. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*.

Oliva, A., and Torralba, A. 2006. Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research*.

Russell, B. C., Sivic, J., Ponce, J., and Dessales, H. 2011. Automatic alignment of paintings and photographs depicting a 3d scene. In *3D Representation and Recognition (3dRR)*.

Schodl, A., Szeliski, R., Salesin, D. H., and Essa, I. 2000. Video textures. In *SIGGRAPH*.

Shechtman, E., and Irani, M. 2007. Matching local self-similarities across images and videos. In *CVPR*.

Sivic, J., and Zisserman, A. 2003. Video google: A text retrieval approach to object matching in videos. In *ICCV*.

Snavely, N., Garg, R., Seitz, S. M., and Szeliski, R. 2008. Finding paths through the world's photos. *ACM Transactions on Graphics*.

Tieu, K., and Viola, P. 2004. Boosting image retrieval. *IJCV*.

Torralba, A., Fergus, R., and Freeman, W. T. 2008. 80 million tiny images: a large database for non-parametric object and scene recognition. *IEEE PAMI*.

Wexler, Y., Shechtman, E., and Irani, M. Space-time completion of video. *IEEE PAMI*.

Whyte, O., Sivic, J., and Zisserman, A. 2009. Get out of my picture! internet-based inpainting. In *BMVC*.

Wolf, L., Hassner, T., and Taigman, Y. 2009. The one-shot similarity kernel. In *ICCV*.