Cross-domain image matching improved by visual attention

Ernani Viriato de Melo
Federal Institute of Triângulo Mineiro
Rua Edilson Lamartine Mendes 300
38045-000, Uberaba, Minas Gerais, Brazil
ernanimelo@iftm.edu.br

Sandra de Amo
School of Computing
Federal University of Uberlândia
Av. João Naves de Ávila 2121 Bloco 1B
38400-902, Uberlândia, Minas Gerais, Brazil
deamo@ufu.br

Denise Guliato
School of Computing
Federal University of Uberlândia
Av. João Naves de Ávila 2121 Bloco 1B
38400-902, Uberlândia, Minas Gerais, Brazil
guliato@ufu.br

ABSTRACT

A good accuracy in image retrieval across different visual domains, such as photos taken over different seasons or lighting conditions, paintings, drawings, hand-drawn sketches, still is a big challenge. This paper proposes the use of visual attention to estimate the relative importance of some regions in a given query image. Recently, researchers used different databases in specific domains to validate their hypothesis. In this paper, we also propose a database with multiple image domains, called UFU-DDD. We used the UFU-DDD database to demonstrate the performance and accuracy gains from the association of visual attention with orientation-based feature descriptors. The analysis of the results showed that our approach outperforms all the standard descriptors used in the experiments. We hope the UFU-DDD database constitutes a valuable benchmark to the future research in cross-domain similarity searching.

Keywords

Visual attention, image matching, saliency, image retrieval in cross domain, painting, sketches.

1. INTRODUCTION

With ever-faster computers and internet connection, the acquisition of collections of images and videos has become an action of our daily lives. Multiple images may possess exactly the same content across a wide range of visual domains, e.g., photos, paintings, sketches, computer-generated images (CG images), with dramatic variations in lighting conditions, seasons, ages, and rendering styles. The development of methods to efficiently compute the visual similarity between images in different domains is a challenge and an urgent need for various applications, such as scene completion [Hay07], Sketch2Photo [Cao11, Eit11], Internet re-photography [Shr11], painting2GPS [Shr11], and CG2Real [Joh11]. Figure 1 illustrates an example of an application where the user gives a painting of the Coliseum as the query and wants to retrieve photos, paintings, sketches and drawings from the same tourist spot.

The task of comparing images in different domains is very challenging, because small perceptual differences can result in arbitrarily large differences at the raw pixel level. In addition, it is very difficult to develop a generalized solution for multiple potential visual domains. For this task, it is necessary to capture the important visual structures that make two images appear similar. Several different image descriptors have been proposed in the literature based on color, shape and texture. In particular, aiming at representing the salient regions (i.e., high gradient and high contrast) of the image, some descriptors have been proposed in the state of art, such as: SIFT - Scale Invariant Feature Transform [Low99], GIST [Oli06] and HOG - Histogram of Gradients [Dal05].

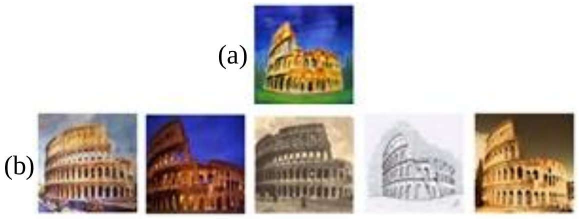

Figure 1. An example of a desired answers list in a cross-domain database: (a) the query; (b) the top-5 answers list of Coliseum in different visual domains.

Recently, researchers have made significant progress in the study of visual similarity in different domains, such as data-driven uniqueness paradigm proposed by Shrivastava et al. [Shr11]. This paradigm aims at focusing on the most important visual parts of the query image. The central idea is to identify the parts of the image are more unique or rare. To that end, Shrivastava et al. [Shr11] proposed to train a linear classifier SVM (Support Vector Machine) at query time, using one feature descriptor with the aim of identifying the uniqueness parts of the image. This method was modified in [Sun13], with an approach that uses multiple features for training the classifier at query time. Although promising results have been achieved, the computational cost to train a SVM classifier for each query is extremely high, preventing its application in real time problems.

A promising alternative to identify relevant parts of an image, with low computational cost, can be found in the studies being conducted on the psychology field related to visual attention. Its central idea is that the most important regions of the image are those that most attracts people's attention. Then, the features extracted from these regions may be strongly weighted for the image retrieval task. Several works [Bor09, Sat10, Soa12] use visual attention to identify different ways to obtain regions of interest and are focused on other tasks such as classification, separation of foreground and background, object recognition, image retrieval. To the best of our knowledge, there is no reference in the literature addressing the use of visual attention in the context of cross-domain image retrieval.

The main goal of this paper is to show that visual attention models can identify the relevant parts of the query image and when associated with image descriptors can contribute to the improvement of the similarity searching accuracy in different visual domains with low computational cost. Differently from the strategy proposed in [Shr11, Sun13], our approach can be computed in real time. An example is illustrated in Figure 2.

The main contributions of this paper can be summarized as follows:

1. We built a new database with images in different visual domains, called UFU-DDD. The database contains 22 classes and each class is composed of 50 images of the same scene in different visual domains. Some examples of images in our database are showed in Figure 4. To our best knowledge, this database is the first one that put together, in a same images class, scenes obtained from several different visual domains, such as photos took over different seasons or lighting conditions, paintings, drawings, computer graphic (CG) images, and sketches. To date, the state-of-the-art databases are constructed to evaluate methods to perform the matching between specific domains [Cho08, Eit11, Rus11] and not for multiple visual domains as proposed in this work. Sometimes, the databases are dynamically created for each query, limiting the comparison and the importance of the experiments [Shr11, Sun13].
2. We proposed a new strategy to associate visual attention maps with well-known orientation based image descriptors such as SIFT [Low99], GIST [Oli06] and HOG [Dal05]. The results showed that our approach overcomes the conventional ones in cross-domain image retrieval.

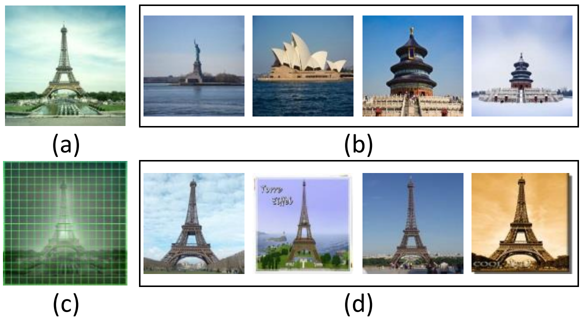

Figure 2. An example of the use of visual attention in an image retrieval task: (a) a photography of the Tower Eiffel; (b) the top-4 answers for the image query in (a) without the use of visual attention, models; (c) the visual attention map of the query image, superimposed by a regular grid. Note that the tower region is now highlighted with respect to the image background; (d) the top-4 answers for the query using an association of the visual attention map with image descriptors in the image retrieval process.

The remainder of this paper is organized as follows. We first give an overview of the related work in Section 2. Section 3 describes our proposal to associate visual attention with orientation-based feature descriptors. The methodology of the experiment and the analysis of the results are provided in Sections 4 and 5 respectively. We discuss about the limitations of our approach and future work in Section 6.

2. BACKGROUND REVIEW

We briefly review related works on cross-domain matching, orientation-based feature descriptors, and models of visual attention.

2.1 Cross-domain Matching

A place, scene, or objects can be recorded in an image in different visual forms, which we call visual domains. Nowadays it is common to find databases containing images with the same semantic content, but in different domains, such as photographs taken over different seasons or lighting conditions, paintings, sketches, drawings, CG images, etc. Many studies have been dedicated to match images between specific domains, such as photos under different lighting conditions [Cho08], sketches to photographs [Cao11, Eit11], paintings to photographs [Rus11], and CG images to photographs [Joh11]. However, these specific domain solutions are not directly extensible to multiple domains.

For a generalized solution for this problem, we highlight three works. In the first one, proposed by Shechtman and Irani in [She07], the authors describe an image in terms of local self-similarity descriptors (SSIM) that are invariant for cross visual domains applications. The second one proposed by Shrivastava et al. [Shr11] and third one proposed by Sun et al. [Sun13] are based on the data-driven uniqueness paradigm. Shrivastava et al. proposed to train a linear Support Vector Machine (SVM) classifier for each query, in query time, to set weights to each dimension of either the HOG or SIFT descriptors. Based on this same idea, Sun et al. proposed to use weighted vectors for multiple features (Filter Bank, SSIM and HOG simultaneously) with weights associated with the each dimensions of these descriptors. The time spent to run a query on a database retrieval of 5,000 images in the method proposed in [Shr11] is under around three minutes on a 200-node cluster, while in the method proposed in [Sun13] is greater than 10 minutes on a PC with a 3.40 GHz Intel i7 CPU and 8 GB RAM. Although promising results have been achieved in both the solutions, these strategies presented a high computational cost, preventing their application in real time similarity searching.

2.2 Orientation-based Descriptors

The descriptors extractors are methods to derive automatically visual information from an image and organize them into a feature vector that represents the image content. In image retrieval run in different visual domains, the locally salient parts of the image are highly relevant information in the calculation of visual similarity. With this aim, several descriptors have been presented in the literature, among them are: SIFT, GIST, and HOG.

2.2.1 Spatial Pyramid SIFT Descriptor

Scale Invariant Feature Transform – SIFT, proposed by Lowe [Low99], is a descriptor that detects a set of keypoints and describe a neighbor of each one in terms of the frequency of gradient orientation. The result is a 128-dimensional vector to describe each keypoint. The SIFT descriptor is invariant to translation, rotation, scale, and illumination conditions.

With the aim of addressing the similar image retrieval images of objects taken in different views, several works use the Bag of Visual Words (BoVW) [Siv03, Csu04, Laz09, Soa12]. In a general way, the BoVW consists of identifying, sparsely or densely, a set keypoints, in a training image database, and cluster them in a predefined number of groups. Each group is referred as a visual word. Then, all the image in the database and the query image are represented by a frequency histogram of visual words. Two images are said similar if their histograms are close to each other according to a similarity measure.

In this paper, we are particularly interested in an extension for the BoVW proposed by Lazebnik, Schmid and Ponce [Laz09] termed Spatial Pyramids (SP). They use a dense regular grid to detect the keypoints. Firstly, SIFT descriptors of *16* x *16* pixel patches of all image database are computed over a grid with spacing of *n* pixels. Then, using the k-means algorithm, SIFT descriptors are grouped and the representative of each group are used to build the visual words dictionary. The frequency histogram of visual words is computed for each image, as in BoVW. Now, the spatial pyramid is computed by partitioning the image into regular sub-regions in several levels, and assessing a frequency histogram for each sub-region. The process continues until it reaches a predetermined number of pyramid levels. The final descriptor is composed by concatenating all frequency histograms derived in the process.

2.2.2 GIST Descriptor

The GIST descriptor presents good results in scene categorization and image retrieval [Oli06]. The idea is to develop a statistical representation with low dimensionality of the scene. The GIST descriptor computes the energy using a bank of Gabor-like filters evaluated at all orientations and different scales for each of the cells obtained by chopping up the image into *N* by *N* pieces. The format of the Gist descriptor is a vector with [*scales*] * [*orientations*] * [*number of cells*] dimensions.

2.2.3 HOG Descriptor

The Histogram of Oriented Gradients – HOG – descriptor [Dal05] was firstly proposed to deal with human detection task and later became a very popular feature in object detection area. When extracting HOG features, the orientations of gradients are usually quantized into histogram bins and each bin has an orientation range. An image is divided into overlapping cells and in each cell a histogram of oriented gradients falling into each bin is computed and then normalized to overcome illumination variation problems. The features extracted from all the cells are then concatenated together to form the HOG descriptor of the whole image, with [*cells*] * [*bins*] dimensions. The HOG features are similar to SIFT descriptor, but HOG features are computed in dense grids at some single scale without orientation alignment. In this paper, we used the HOG descriptor algorithm proposed in [Fel10], which uses 31 orientations bins to compute the histogram of oriented gradients. The similarity between the image query and other image can be computed by various functions of distance, for example, the Canberra distance.

2.3 Models of Visual Attention

Humans are faced with an overwhelming amount of visual information. However, this amount of information is much larger than all the information that the brain processes and assimilates. By rapid eye movements, referred to as saccades movements, the brain must prioritize and receive only part of the visual information at every instant.

Visual attention is the ability of the human visual system to select and process only the most important regions in a scene, while ignoring the rest of the image information. Intuitively, saliency characterizes some parts of a scene which appear with high relevance for an observer. A saliency map indicates the conspicuity of each pixel of the scene, i.e. the probability of parts of the scene to attract the attention of humans. The saliency map is visualized as a grayscale image, where the brightness of a pixel is proportional to its salience. Models of visual attention try to represent the mechanism of visual attention by saliency map. A nice survey about saliency map is presented by Borji and Itti in [Bor13]. The authors presented a classification of attention models into seven categories, considering the strategy used to obtain the saliency map. We evaluate one model of each one category and the GBVS model, proposed by Harel et al. [Har07], presented the best results for the cross-domain problem. Harel et al. proposed a bottom-up saliency map which uses the Markov chains over various feature maps and treats the equilibrium distribution over map locations as activation and saliency values. They proposed to unify the activation map and normalization/ combination maps steps by using dissimilarity and saliency to define edges on graphs which are interpreted as Markov chain. In this work, we'll use the GBVS model in the experiments.

3. OUR APPROACH

In this paper we propose combining visual attention models with image descriptors for image retrieval in different visual domains. Our hypothesis is that the relevant regions of an image, highlighted by a saliency map, are more important to characterize an image for content based image retrieval. Figure 3 summarizes the idea proposed in this paper.

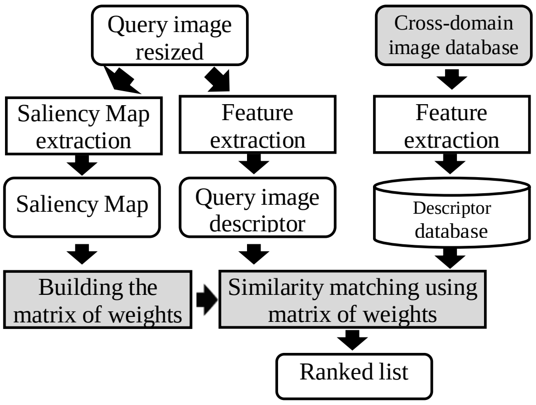

Figure 3. The flowchart of the association of visual attention with image descriptors. The darker boxes highlight our contributions.

Following Figure 3, for the query image, two processes occur in parallel: saliency map extraction, and feature extraction. The saliency map extraction can be performed by any of the visual attention models already developed. Because the nature of cross domain database, we will use the feature extraction method based on gradient / contrast orientation, that are SIFT, GIST and HOG descriptors, but the image descriptors are not limited to them and depend on the application.

After extracting the saliency map, the saliency magnitude is normalized to the range [0, 1], and then it is superimposed by an *NxN* size grid. After that, the matrix of weights, with *NxN* dimension, is derived with the aim of highlighting the relevant parts of the image while discard or attenuate the importance of the remainder regions. To that end, the value of the position (*i,j*) of the matrix of weights is assessed by computing the median saliency magnitude value of the normalized saliency map in the corresponding grid cell. Those median values that are lower than a given threshold *T* is set zero, indicating the low importance of that image for the representation of the scene. The matrix of weights is built as showed in Algorithm 1. Finally, we use information of the matrix of weights in same way to run the similarity searching.

Algorithm 1: Building of matrix of weights

```
MatrixOfWeights(I, MV, T, N)
  Input: I: image; MV: model of visual attention;
         T: threshold; N: grid size
  Output: W: N x N matrix of weights
begin
  SM = SaliencyMapExtraction(I, MV)
  SM = Normalize(SM, 0, 1)
  for each par(i, j) Є grid do
      Md = median(SM, i, j)
      if (Md < T)
         W[i,j] = 0
      else
         W[i,j] = Md
  return W
End
```

3.1 Visual Attention with Spatial Pyramid/ SIFT (VA-SP-SIFT)

In our approach, only the SIFT descriptors extracted from the relevant regions are used to construct the Spatial Pyramid, as shown in Algorithm 2. The similarity measure between the image query and other images is computed by the histogram intersection function, as described in [Laz09].

3.2 Visual Attention with GIST (VA-GIST)

The matrix of weights, different from the strategy described in Section 3.1, is used to weight the similarity measure in query time.

Let I and J be the query image and the target image. Let W^I, G^I and G^J be the matrix of weights of I, obtained as described in Algorithm 1, the GIST vector for I and the GIST vector for J, respectively. $G(c,s,r)$ is GIST vector value of the cell c in a scale s and orientation r, where $c = \{1, \ldots, m\}$, $m = NxN$ is a number of cells, $s = \{1, .., p\}$, p is a number of scales, $r = \{1, \ldots, q\}$, q is the number of gradient orientations. In this work we propose to use the Weighted Euclidean distance as defined in Eq. 1, to determine the similarity between I and J.

$$D(I,J): \sqrt[2]{\sum_{c=1}^{m}\sum_{s=1}^{p}\sum_{r=1}^{q}\left(1-\left(G^I(c,s,r)-G^J(c,s,r)\right)^2\right) W^I(c)} \quad (1)$$

where $W^I(c)$ is a weight associated with the c^{th} cell in the matrix of weights.

Algorithm 2: VA-SP-SIFT Descriptor

```
Data: I: image query; MV: model of visual attention;
      T: threshold; N: grid size;     D: dictionary
Result: V: VA-SP-SIFT descriptors vector
begin
    W = MatrixOfWeights (I, MV, T, N)
    SIFTS = Dense-SIFT(I)   //as in [Laz09]
    SIFTS_VA = Fusion (SIFTS, W) //only the SIFT
                  descriptors that fall within the cells
                  where the corresponding position in W
                  is different from zero are kept.
    V = ProduceSP(SIFTS_VA, D) //as in [Laz09]
    return V
end
```

3.3 Visual Attention with HOG (VA-HOG)

For HOG, the proposal is similar to that one described in Section 3.2.

Let I and J be the query image and the target image. Let W^I, H^I and H^J be the matrix of weights of I, see Algorithm 1, the HOG vector for I and the HOG vector J, respectively. $H(c,i)$ is the normalized count of the i^{th} orientation bin of the c^{th} cell, where $c = \{1,..,m\}$, m is a number of cells, $i = \{1, .., 31\}$. $W^I(c)$ is a weight associated to c^{th} cell. The similarity between I and J is computed by using the Weighted Canberra distance, as shown in Eq. 2.

$$D(I,J) = \sum_{c=1}^{m}\left(\sum_{i=1}^{31} 1 - \frac{|H^I(c,i) - H^J(c,i)|}{|H^I(c,i)| + |H^J(c,i)|}\right) W^I(c) \qquad (2)$$

4. METHODOLOGY

We run several experiments in order to analyze the performance of the use of visual attention for cross-domain image matching. The experiments are divided according to the feature descriptor used. The results of quantitative analysis are reported in terms of the Average Precision (AP) values at the top-k answers.

4.1 Databases

The scientific community needs a unique database to evaluate methods towards image retrieval across visual domains. Aiming to address this need, we created a public database, called UFU-DDD (Database of Different domains of University of Uberlândia). We also use 10,000 images from the MIRFLICKR Database [Hui08] to test the robustness of our proposal.

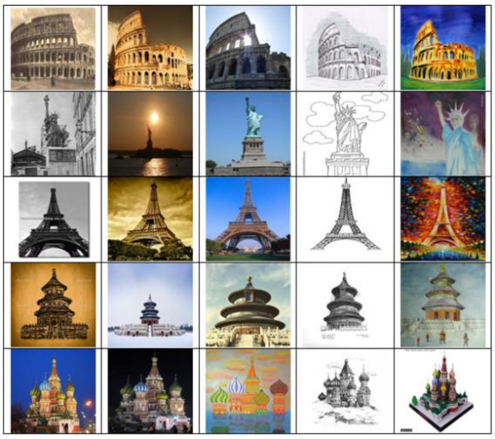

Figure 4. A sub-set of classes of UFU-DDD: Coliseum, Statue of Liberty, Eiffel, Temple of Heaven, Saint Basil's Cathedral.

Database of Different Domains of University of Uberlândia (UFU – DDD) – we have created a new database comprised of 1,100 images. The UFU-DDD database was collected by crawling images from google images website using keywords about tourist spot such as "painting of Tower Eiffel", "sketch of Cathedral San Basilio". This procedure was necessary because we did not find a database with the particularities that we consider important to evaluate our approach. In order to obtain classes with a variety of domains, we decided that each class would be a tourist spot with exactly 50 images across different domains, totaling 22 classes. The tourist spots are many, such as: waterfall, church, coliseum, temple, stadium, castle, museum, opera house, etc. The database contains 91 very old photographs, 677 photographs under different lighting and stations, 150 sketches, drawings and CG images, and 182 paintings. In all the cases the foreground is centered in the image. Figure 4 shows five classes of UFU-DDD and each class with images in different visual domains. With UFU-DDD it is possible to design experiments such as: given a query painting, what are the photographs, drawings, and sketches more visually similar? Given an old photograph, what are the recent photographs of the same place? Given a sketch, which are the paintings of the corresponding place?

MIRFLICKR Database (MIF) – this database contain 1 million Flickr images under the Creative Commons license. It is commonly used for the visual concept detection, photo annotation and image retrieval task. We used 10,000 photographs of this database just to test the performance of our approach under different database sizes.

4.2 Experiment Setup

In all experiments, the images were resized to 200 x 200 pixels. The extraction of the saliency map was done using GBVS model proposed by Harel et al. [Har07]. For each image descriptor we compared our approach against a publicly available, third-party authored reference, implemented in Matlab. The parameters used to derive each descriptor are described following.

VA-SP-SIFT: for each image, we compute spatial pyramid representation with 3 pyramid levels using Dense-SIFT descriptors of 16x16 pixels patches computed over a grid with spacing of 8 pixels. We used a vocabulary of 400 visual words. After several tests, we empirically set the threshold value at 0.2 according to Algorithm 1.

VA-GIST: we computed the GIST representation for each image using an 8 x 8 grid, 4 scales, and 8 orientations. The threshold T in Algorithm 1 is set at zero. All the weights are used in the Weighted Euclidean distance, as defined in Eq. 1.

VA-Normalized-HOG (VA-NHOG): for each image, we compute HOG representation with 625 cells of 8 x 8 pixels divided in a 25 x 25 grid, and 31 orientations bins. We experimented both implementation, the standard HOG descriptor as well a simple normalized HOG (NHOG). The NHOG vector (VNHOG) is defined as a zero-centered version of HOG vector (VHOG), where VNHOG = VHOG – mean(VHOG). We perceived slightly better results with NHOG and then we adopted it in our experiments. The threshold T in Algorithm 1 is empirically set at 0.3. We also evaluate different similarity measures, such as Cosine, Chi-square, Euclidean, Manhattan, Canberra distances. We adopted the Weighted Canberra distance, as defined in Eq. 2.

5. EXPERIMENTS

In order to evaluate the performance of our approach, we conducted three experiments, each one using a specific domain as the query image. The experiments are run using the UFU-DDD database. We also used images from the MIRFLICKR database as distractors, in three different versions: UFU-DDD + 3,000 MIF; UFU-DDD + 6,000 MIF; and UFU-DDD + 10,000 MIF.

5.1 Photograph as Queries

In this experiment, the query images are photos took over different ages, seasons, weather or lighting conditions. We collected from the UFU-DDD a dataset of 44 photos (2 photos of each class) to be used as queries. Table 1 shows the Average Precision at top 10 (AP@10) and top 30 (AP@30). In all the cases, our approach obtained an important improvement in the results when compared to the standard descriptors. The gain obtained for our proposal varies from 9% to 30% for top 10 and from 5% to 15% for top 30. The gain depends on the descriptor and the database size. It is worth noting that the inclusion of distractors in the database did not affect the gain obtained with the use of visual attention. Figure 5 (a) shows the top 3 answers for the Sydney Opera House with and without the use of visual attention. This example illustrates the superiority of our approach.

Methods	Databases							
	UFU-DDD		UFU-DDD+ 3,000MIF		UFU-DDD+ 6,000MIF		UFU-DDD+ 10,000MIF	
	@10	@30	@10	@30	@10	@30	@10	@30
SP-SIFT	0.46	0.29	0.40	0.24	0.38	0.22	0.31	0.20
VA-SP-SIFT	**0.64**	**0.39**	**0.51**	**0.31**	**0.47**	**0.27**	**0.43**	**0.25**
GIST	0.58	0.36	0.55	0.32	0.53	0.31	0.52	0.29
VA-GIST	**0.79**	**0.46**	**0.75**	**0.42**	**0.74**	**0.40**	**0.72**	**0.38**
NHOG	0.53	0.35	0.46	0.29	0.43	0.27	0.40	0.24
VA-NHOG	**0.83**	**0.50**	**0.71**	**0.40**	**0.66**	**0.36**	**0.63**	**0.33**

Table 1. Average Precision at top 10 and top 30 with different database sizes. Queries: Photo.

Methods	Databases							
	UFU-DDD		UFU-DDD+ 3,000MIF		UFU-DDD+ 6,000MIF		UFU-DDD+ 10,000MIF	
	@10	@30	@10	@30	@10	@30	@10	@30
SP-SIFT	0.31	0.20	0.28	0.16	0.26	0.14	0.25	0.13
VA-SP-SIFT	**0.43**	**0.25**	**0.36**	**0.20**	**0.32**	**0.19**	**0.31**	**0.17**
GIST	0.43	0.26	0.38	0.21	0.36	0.20	0.34	0.18
VA-GIST	**0.57**	**0.34**	**0.51**	**0.30**	**0.47**	**0.27**	**0.46**	**0.26**
NHOG	0.41	0.26	0.36	0.20	0.33	0.18	0.31	0.16
VA-NHOG	**0.65**	**0.41**	**0.53**	**0.32**	**0.49**	**0.29**	**0.46**	**0.27**

Table 2. Average Precision at top 10 and top 30 with different database sizes. Queries: Sketch / Drawing.

5.2 Sketch/Drawing as Queries

Here, the query images are sketches and drawings. We collected a dataset of 44 sketches and drawings (0 to 3 images of each class) to be used as queries (two classes do not contain sketches). Matching sketches/drawings to real scenes is a difficult task. The sketches and drawings are abstract and show strong local deformations with respect to the real scene. Table 2 shows the AP@10 and AP@30. In all cases, it is possible to note an improvement in the results that vary from 6% to 24% for top 10 and from 4% to 15% for top 30. Figure 5 (b) shows a qualitative examples corresponding to the top 3 answers for each descriptor, using or not the attention model. It can be seen that our approach returned 3 relevant photos as answers for sketch used as query. The 3 images returned are relevant to the 3 feature descriptors.

5.3 Painting as Queries

We collected a dataset of 44 paintings (1 to 3 images of each class) to be used as queries. Matching paintings to scenes is also a difficult task because: i) the presence of strong local gradients due to brush strokes; and ii) the painting styles may vary from painter to painter. Table 3 shows the AP@10 and AP@30. In all cases, our approach outperforms the standard descriptors. The gain reached by our proposal varies from 6% to 17% for top 10 and from 3% to 9% for top 30. A Qualitative example is showed in Figure 5 (c). Due to the difficult of match painting to photos or sketches, the standard descriptors failed in all the answers while our approach returned at least two relevant answers in three.

Methods	Databases							
	UFU-DDD		UFU-DDD+ 3,000MIF		UFU-DDD+ 6,000MIF		UFU-DDD+ 10,000MIF	
	@10	@30	@10	@30	@10	@30	@10	@30
SP-SIFT	0.38	0.24	0.31	0.18	0.29	0.16	0.26	0.14
VA-SP-SIFT	**0.50**	**0.31**	**0.39**	**0.24**	**0.35**	**0.20**	**0.32**	**0.17**
GIST	0.44	0.26	0.40	0.22	0.37	0.20	0.35	0.19
VA-GIST	**0.58**	**0.34**	**0.53**	**0.29**	**0.49**	**0.27**	**0.46**	**0.26**
NHOG	0.43	0.28	0.36	0.20	0.31	0.16	0.28	0.14
VA-NHOG	**0.60**	**0.37**	**0.48**	**0.28**	**0.44**	**0.24**	**0.40**	**0.21**

Table 3. Average Precision at top 10 and top 30 with different database sizes. Queries: Painting.

6. CONCLUSIONS

In this paper, we presented a new strategy with low computational cost to highlight the most important parts of an image query with the purpose of images retrieval in databases that contain images in different visual domains. The strategy was evaluated with a different database sizes. We showed that our approach outperforms the standard descriptors. However, this strategy is strongly dependent on the model of visual attention to be used. A typical failure is showed in Figure 6. In this example, our approach fails to find good top matches because the attention model was not able to identify all the body of the Statue of Liberty. Further works are in progress to detect relevant parts of an image, interactively by using eye tracker device.

7. ACKNOWLEDGMENTS

We thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG), and Conselho de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the financial support.

8. REFERENCES

[Bor09] Borba, G. B., Gamba, H. R., Marques, O., and Mayron, L. M. Extraction of salient regions of interest using visual attention models. In IS&T/SPIE Electronic Imaging, 2009.

[Bor13] Borji, A., and Itti, L. State-of-the-Art in Visual Attention Modeling. TPAMI, 35(1): 185-207, 2013.

[Cao11] Cao, Y., Wang, C., Zhang, L., and Zhang, L. Edgel index for large scale sketch-based image search. CVPR, pp. 761–768, 2011.

[Cho08] Chong, H. Y., Gortler, S. J., and Zickler, T. A perception-based color space for illumination-invariant image processing. ACM Trans. Graph. 27(3), 61, 2008.

[Csu04] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. Visual categorization with bags of keypoints. ECCV, 2004.

[Dal05] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection.CVPR, 2005.

[Eit11] Eitz, M., Hildebrand, K. , Boubekeur, T. and Alexa, M. Sketch-based image retrieval: benchmark and bag-of-features descriptors. TVCG, 17(11), 1624–1636, 2011.

[Fel10] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. TPAMI, 32(9), 2010.

[Har07] Harel, J., Koch, C., and Perona, P. Graph-based visual saliency. Advances in Neural Information Processing Systems, 19, 2007.

[Hay07] Hays, J. and Efros, A. A. Scene completion using millions of photographs. In ACM Trans. Graph. 26(3), 2007.

[Hui08] Huiskes, M. J., and Lew, M. S. The MIR Flickr Retrieval Evaluation. ACM MIR'08, Vancouver, Canada, 2008.

[Joh11] Johnson, M. K., Dale, K., Avidan, S., Pfister, H., Freeman, W. T., and Matusik, W. CG2Real: improving the realism of computer generated images using a large collection of photographs. TVCG. 17(9), 1273–1285, 2011.

[Laz09] Lazebnik, S., Schmid, C., and Ponce, J. Spatial pyramid matching. In Object Categorization: Computer and Human Vision Perspectives. Cambridge University Press, 2009.

[Low99] Lowe, D. G. Object Recognition from local scale–invariant features. ICCV, 1999.

[Oli06] Oliva, A., and Torralba, A. Building the gist of a scene: the role of global image features in recognition. Progress in Brain Research. 155, 23-36, 2006.

[Rus11] Russell, B. C., Sivic, J., Ponce, J., and Dessales, H. Automatic alignment of paintings and photographs depicting a 3D scene. 3dRR, 2011.

[Sat10] Sato, M., and Katto, J. Performance improvement of generic object recognition by using seam carving and saliency map. IWAIT, 2010.

[She07] Shechtman, E., and Irani, M. Matching local self-similarities across images and videos. CVPR, 1–8, 2007.

[Shr11] Shrivastava, A., Malisiewicz, T., Gupta, A., and Efros, A. A. Datadriven visual similarity for cross-domain image matching. ACM Trans. Graph. 30(6), 154, 2011.

[Siv03] Sivic, J., and Zisserman, A. Video google: A text retrieval approach to object matching in videos. IEEE Int. Conf. on Computer Vision,. 1470-1477, 2003.

[Soa12] Soares, R. C., Silva, I. R., and Guliato, D. Spatial Locality Weighting of Features using Saliency Map with a Bag-of-Visual-Features. ICTAI, 1-6, 2012.

[Sun13] Sun, G., Wang, S., Liu, X., Huang, Q., Chen, Y., and Wu, E. Accurate and efficient cross-domain visual matching leveraging multiple feature representations. The Visual Computer, 29, 565-575, 2013.

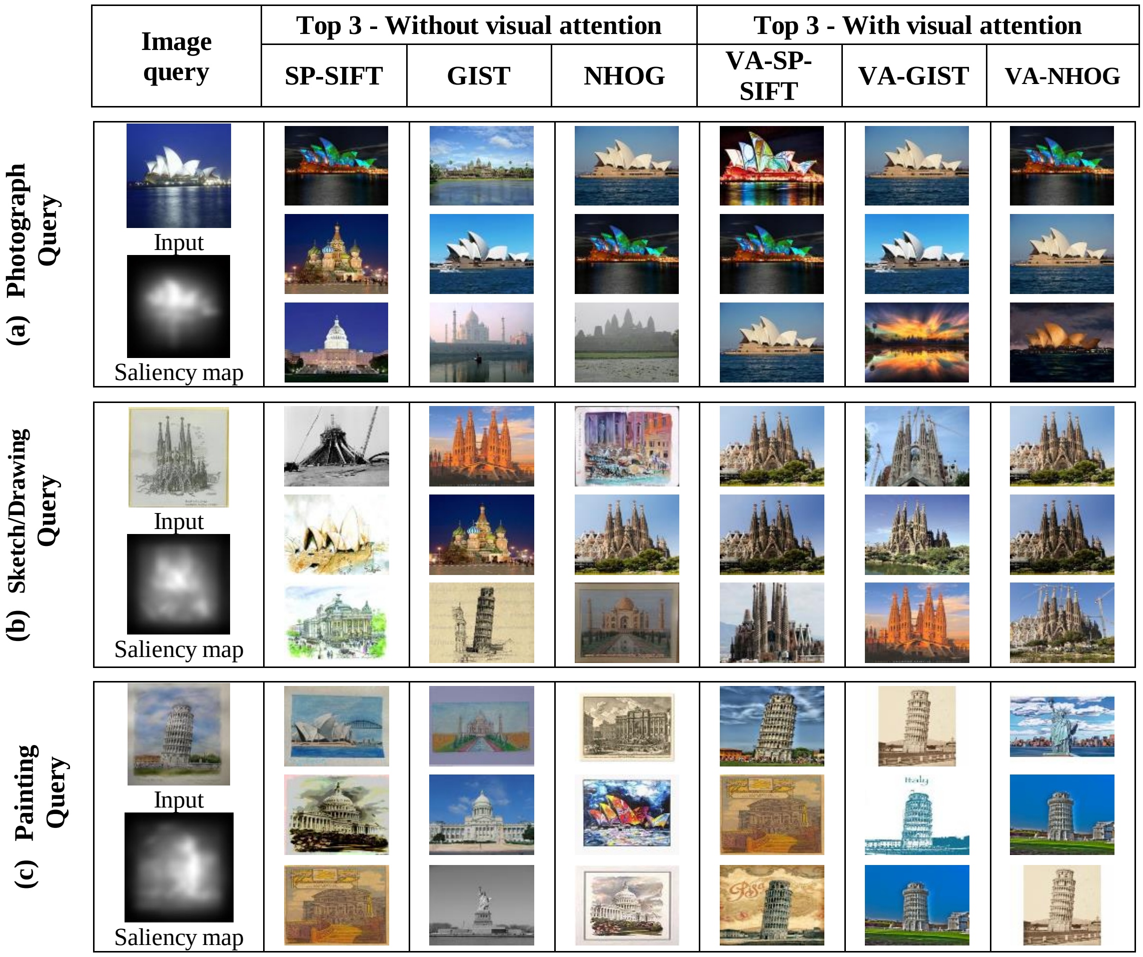

Figure 5. Qualitative comparison of feature descriptors with and without the GBVS model.

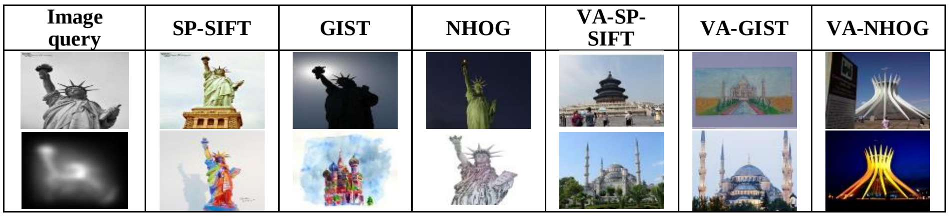

Figure 6. Typical failure case. The GBVS model mainly highlighted the head and torch of the Statue of Liberty as relevant regions. However, regions as the body and the sky around the Statue of Liberty are important for CBIR.