CONTENT-BASED RETRIEVAL OF IMAGES FOR CULTURAL INSTITUTIONS USING LOCAL DESCRIPTORS

Eduardo Valle, Matthieu Cord, Sylvie Philipp-Foliguet

Équipes Traitement des Images et du Signal - ETIS UMR CNRS 8051
6, ave du Ponceau, 95014 Cergy-Pontoise, France

ABSTRACT

The task of identifying an image whose metadata are missing is often demanded from cultural image collections holders, such as museums and archives. The query image may present distortions (cropping, rescaling rotations, colour changes, noise...) from the original, which poses an additional complication. The majority of proposed solutions are based on classic image signatures, such as the colour histogram. Our approach, however, follows computer vision methods, and is based on local descriptors. In this paper we describe our approach, explain the SIFT method on which it is based and compared it to the Multiscale-CCV, an established scheme employed in a large scale practical system. We demonstrate experimentally the efficacy of our approach, which achieved a 99.2% success rate, against 61.0% for the Multiscale-CCV, in a database of photos, drawings and paintings.

1. INTRODUCTION

Institutions possessing large image databases, such as museums, archives, and news agencies, often face the separation of an image from its metadata: the visual document is present, but its title, authors, description and other relevant information are missing. This arises because source references are frequently absent or irregular. Because the meaning of a document is dependent on its context, the lack of metadata reduces its usefulness, and the institutions are often asked to retrieve all information about a document, using only the visual information as query.

The matter is made worse because the query image may have been distorted from the original:
- Geometric transforms: mainly translations, rotations and scale changes.
- Colorimetric transforms: changes in brightness, contrast, saturation or colour balance.
- Selection and occlusions: the image may have been cropped to select just a detail; it may also present labels, annotations, censor bands, etc.
- Compression: images are often stored with lossy compression schemes, like JPEG.
- Other distortions: the image may be noisy, fuzzy, dithered (for offset printing), etc.

The use of watermarking techniques to embed a unique ID in each original image may alleviate the problem, but there are serious shortcomings: first, it assumes all reproductions come from the same digital source; second it is not applicable to the images reproduced before the adoption of the watermarking; third, watermarking techniques are unreliable under strong image distortions.

Many indexing techniques have been proposed to retrieve images without metadata [7]. Those methods must provide at least region level description; otherwise their performance would be unacceptable in the presence of occlusions or cropping [1]. Dealing with strong contrast and luminance changes is problematic to many colour/texture based methods. Colour based methods perform poorly under colour balance transformations and fail completely under greyscale conversions. Cropping, occlusions, colour shifting and conversion to greyscale are fairly common in the image management industry, which leads to dissatisfaction with traditional methods.

Local descriptors have been proposed to solve several problems in computer vision, from point matching in stereovision, to object detection [10]. They are very robust to occlusions, cropping and geometric transformations.

We propose in this paper to adapt one of the most popular computer vision approaches — the SIFT points of interest [2, 6] — to our context: Content Based Information Retrieval for cultural image collections. We also evaluate the performance of our approach, comparing it to a good archetype of the time-honored indexing methods, the Color Coherence Vector [3], which was successfully implemented in a large scale European software project [11].

2. THE BASELINE METHOD — CCV

Because colour is so distinguishing, colour histograms have been used for image indexing since the establishment of this discipline [9]. Unfortunately, histograms discard all data about spatial distribution of colour, reducing the signature efficiency.
A relatively simple attempt to preserve some of the spatial information is the Colour Coherence Vector or CCV. The CCV can be viewed as a double colour histogram, where the frequency of each colour is accounted in two separated entries: one for large uniform regions (called coherent regions) and one for small, insulated regions [3].

Figure 1. This sample image shows a coherent region (a) and an insulated region (b) for the colour black

In the original CCV algorithm (that we replicate), a region is considered small if it corresponds to less than 1% of the entire image. The CCV for [Figure 1] would state:
- Black: 28% coherent; 4% insulated;
- White: 68% coherent; 0% insulated.

The signature, originally global, was adapted to allow the retrieval of an extract of an image. This adaptation, named Multiscale-CCV, consists of applying the CCV on blocks of the image [11]. Those blocks belong to two hierarchical grids, the second displaced half-way from the first, in order to capture all possible regions of interest. [Figure 2]

The actual algorithm is not implemented top-down as described above, but bottom-up. The image is divided in patches of 64 × 64 pixels taken from two grids (the second one displaced 32 pixels from the first). The image is then downsampled to half its resolution and the process is iterated, until the image roughly fits in one patch.

Figure 2. Multiscale-CCV region grids, original (above), and displaced (below)

Colour histograms must map the huge colour space of the source images into a relatively small set of entries. The Multiscale-CCV achieves this in a very rudimentary way: the RGB space is cut into 64 “cubes” of equal volume — which, in practice, is done by keeping the two most significant bits of each colour component.

Though rather simple, the Multiscale-CCV has been successfully implemented and used in large scale application for retrieval of documents from fragments without metadata [11]. A detailed description of the CCV descriptor and a comparison with colour histograms may be found in [3].

3. THE SIFT POINTS OF INTEREST METHOD

All points-of-interest (PoI) methods are based on a common ground: the detection and description of some points of the image, which provide more information than the others.

To detect PoI, one can use many criteria: local contrast, local maximization/minimization of certain functions (Laplacian, gradient, etc.), threshold over a curvature function (Harris, Hessian, etc.).

What is required from the PoI is repeatability: a high fraction of the points should be found at the same locations, even if the image suffers geometric or colorimetric deformations. Once a point is detected, a descriptor must be generated for indexation purposes. Usually, only a small patch around the point is analyzed, generating a local descriptor. If the method is to be robust, the descriptor must be invariant under the concerned deformations.

For our comparison, we have chosen SIFT (Scale Invariant Feature Transform), which is one of the most robust methods under similarity transformations (translations, scale changes and rotations) [2].

The SIFT points are local extremes (minima and maxima) in a scale-space composed by differences of Gaussians of progressively larger standard deviations, as explained below. Their description is based on a normalized histogram of the gradient directions around the point. To achieve rotation invariance, the patch is rotated towards the most frequent direction of the gradient.

The SIFT algorithm is a complex one. Its multiple steps carry out the task of detecting highly repeatable points and giving them a highly discriminating description. At several stages, a threshold or a parameter must be chosen (which is done mostly empirically in [2]). We feel that is important to give a detailed description of the method in order to provide the reader an idea of what is at stake:

1. If in colour, the image is converted to greyscale. To avoid aliasing, the image is pre-treated, by doubling its size and then smoothing it with a gaussian filter.
2. A set of progressively stronger gaussian filters is applied to the image generated in step 1. This generates a series of images $G_1, G_n$. The standard deviations are chosen according to the equation below, where $\sigma_1$ is the standard deviation applied to the first image:

   \[ \sigma_i = \sigma_1 \times 2^{(i-1)/k} \]

3. Adjacent images are subtracted to generate a series of differences of gaussian $D_1, D_n$, e.g., $D_1=G_2-G_1, D_2=G_3-G_2$, etc.
4. PoI candidates are the local extremes in the 3D grid formed by the differences of Gaussians series. A point is extreme if and only if its greyscale value is greater or lesser than all its 26 neighbours. [Figure 3]

![Figure 3. The value of interest point X must be greater or lesser than all its 26 neighbours](image)

5. The scale of the point is determined by the scale of the plan where it was found. Its precise position is computed by interpolation of the scale-space grid.

6. A minimum contrast threshold is applied to filter out low contrast points.

7. Edge responses are eliminated to promote corner points (points of high curvature value), using the ratio of the eigenvalues of the Hessian matrix.

8. A direction is assigned to each surviving point. This is done by choosing the smoothed image $G_i$ closest to the scale of the point, taking the gradient of this image, and computing a histogram of the values around the point. The peak direction is selected. If there are other significant peaks, the point can get multiple directions.

The choice of the difference of gaussian is motivated by the fact it approximates the Laplacian function, which performs very well in the composition of a scale-space. Lowe depicts a clever scheme for computing the Gaussians without the heavy processing cost associated to large standard deviations. [2]

Next, it is necessary to give the points a description invariant under the expected transformations, in order to allow indexation and retrieval. SIFT will proceed as follows:

1. The gradient image of the smoothed image closest to the scale of the point is taken, just like we have done in the step (8) of the detection.

2. A gaussian weighting function is set around the point, with a standard deviation equal to half the size of the descriptor window.

3. The descriptor window is divided in zones. We illustrate here a descriptor with 4 zones ($2 \times 2$), though the real descriptor has 16 zones ($4 \times 4$). [Figure 4]

4. In each zone, we compute a histogram of the gradient values, with 8 orientations. To avoid boundary effects, a trilinear interpolation is used to distribute the values of the samples.

5. The descriptor is composed by all entries of all histograms. This generates a quite large vector (32 in our illustration, 128 in the real implementation).

6. To minimize the effects of illumination changes, the vector is adapted. The affine variations are accounted by normalizing the vector. The saturation effects are reduced by applying a maximum threshold to the gradient magnitudes of the histogram, and renormalizing it.

![Figure 4. SIFT Descriptor. The gradient values around the point are weighted by a gaussian function (dotted circle). Each zone has the histogram computed for 8 gradient orientations](image)

### 4. OUR SYSTEM ARCHITECTURE

#### 4.1. Indexing and Querying the Local Descriptors

When using traditional methods, querying the base consists of finding one set of signatures: the most similar under some criterion. When employing local descriptors, however, each operation is composed of many sub-queries, one for each PoI present in the query image, and the accumulation of the partial answers will determine the final results.

For each query point, one searches for the set of $k$ points in the base (the most similar), and counts a vote for each corresponding image. One expects that, even if many images may receive some votes, only the pertinent ones will have a significant amount of votes.

Because the process consists of several inspections (a query image typically has a few hundred points), it is important to perform the search as fast as possible.

Unfortunately, all known methods to search for the closest points, known as nearest neighbours, have high computational costs on high dimensional spaces. SIFT descriptors, whose dimension is 128, incur in such a penalty, that a naïve sequential search will outperform most methods.

In order to overcome this problem, one must accept an approximate nearest neighbour search, in which one are not guaranteed to find the exact $k$ points most similar to the query, but still expects to have most of the good answers, eventually mingled with other vectors not too dissimilar to the true answers.

From the several methods described in the literature, we have chosen the \textit{KD-tree} with the \textit{best bin first search}
strategy [5], which has been successfully used with SIFT before. The KD-tree is a binary search tree, in which each node chooses a dimension from the space of the points being classified: all points with values less or equal to the node in that particular dimension will be put in the left sub-tree; the other nodes will be put in the right sub-tree, and thus recursively [4]. The best bin first is one possible strategy of traversing the KD-tree, in which we visit first the nodes where it is most probable we will find the nearest matches. Using this technique it is possible to avoid the large costs of the sequential search and still have a reasonable approximation of the $k$ nearest matches.

4.2. Adaptations to the CBIR Context

Computer vision based approaches have to be adjusted, when transposed to our context: the databases are much larger, the images have much higher resolutions and sizes, and most of all, we are concerned with the retrieval of flat, 2D documents and not all the possible perspectives of 3D objects. It was predictable that several adaptations were to be made in order to extract the best performance from a point of interest method.

We had to include a pre-processing of the images before the computation of the PoI. We convert the images to grayscale, normalize the grayscale histograms to achieve similar global contrast among the images, and perform a Gaussian smoothing in order to reduce the JPEG blocks. Without this pre-processing, we had a huge number of PoI, many of which were of spurious, spoiling the results.

The original algorithm includes a criterion of significance, which tries to retain only the matches that are discriminating. This is done by comparing, for every query point, the distance to the most similar target and the second most similar target. If those distances are too close, the match is ignored. Since the underlying hypothesis, that there is only one correct match in the base for every point, is false in our case, we had to eliminate this criterion.

Finally, the original SIFT algorithm uses the Hough transform to remove false matches, because in the application considered (object recognition), the number of outliers is sometimes higher than 95% [2]. Because we do not have to deal with the complexity of perspective deformation, in our case, the number of outliers never falls much below the 50% mark. Instead, we adopted RANSAC [8], which performs very well, and is much more straightforward to parameterize than the Hough transform.

5. EXPERIMENTAL SETUP

To compare the two methods, we used a database consisting of 60 original images, each accompanied of 15 transformations, summing up to 960 images. The final base contained almost nine hundred thousand PoI. The database contained old and modern photos, paintings, drawings and old printed images. The transformations were:

- Two changes in **colour**;
- Dithering;
- Addition of a strong impulsive **noise**;
- Softening;
- Three **rotations**;
- Two **extracts**;
- Stretch + skew;
- Insertion of a small **occlusion** in the image;
- Insertion of a monochromatic **frame** around the image;
- **Incrustation** of the image in another image;
- Aggressive compression (up to 1:100).

A sample of these may be seen in [Figure 5].

![Figure 5. Some of the possible transformations for the images.](image)

The experiment consisted in submitting each one of the transformed images as query, and collecting the 25 most similar answers, accordingly to each method. The answer was considered successful if the original image was among the returned set. The choice of cutting at 25, if somewhat arbitrary, was based on the number of thumbnails a user can view simultaneously with some comfort. The idea is that the user will receive a screen of results and decide, in a few moments, if the correct answer is among them.

6. RESULTS

We compiled the results of the experiments in Table 1. We showed the percentage of successes, classified by type of transformation of the query.
decreasing the quality of the classification. Ideally, the value of \( k \) will be outside the neighbourhood set. Conversely, if the value of \( k \) is too small, some images in the correct category will not be found, because the corresponding points will be outside the neighbourhood set. Conversely, if the value of \( k \) is too large, false positives will proliferate, decreasing the quality of the classification. Ideally, \( k \) should be exactly the number of correct matching points for each query point (which corresponds, roughly, to the number of images in each category). In most practical cases this number may be impossible to determine a priori. We evaluated the results for \( k = 5 \) and \( k = 10 \) nearest neighbours.

The performance is better for \( k = 10 \) than for \( k = 5 \) and would probably be optimal at \( k = 16 \) (the actual number of transformations of each image). Unfortunately, as \( k \) increases, so does the running time. Besides, since we use an approximate nearest neighbours search, a large value of \( k \) implies more imprecision.

For both \( k = 10 \) and \( k = 5 \), our approach performed better than or equally well the other methods, for all the transformations.

CCV performance was particularly poor in the colour transformations and was catastrophic under the dithering, because this transformation completely modifies both the value and the coherence of the colours.

For the sake of reference, we included the performance of a simple colour histogram, using the same colour quantization of the CCV (RGB space, 64 equal “cubes”). The results are inferior to the CCV for most of the transformations. The softening and impulsive noise, are remarkable exceptions, explained by the fact they disturb the coherence of the colours.

Though our approach had a much better performance, its computing times were also much higher. A query for an image takes several minutes (against a few seconds for CCV). However, in our context, users are more concerned with recall than with running times, because when they are not able to find the document automatically, the manual search may take entire weeks.

7. CONCLUSIONS AND PERSPECTIVES

We have adapted the SIFT interest points to thrive in the context of CBIR on cultural image databases. Our adaptations include a pre-processing of the images, a modification of the criterion of retention of significant matches, and a different robust matching algorithm.

We have shown that our approach attains excellent performance, being able to retrieval documents without metadata with a success rate of 99.2%, against 61.0% of CCV — the established method in this context, used in a large scale European project — obtains.

Despite the fact that recall is much more important to the users than the running times, we still want to be able to query large databases in a few seconds. Our current work, on high-performance computation of approximate nearest neighbours, will allow us to be as fast as the traditional methods, and much more accurate. We will be able, then, to apply our approach in other contexts, where timeliness is critical.

8. ACKNOWLEDGEMENTS

Eduardo Valle is sponsored by a CAPES scholarship through the CAPES/COFECUB program. The original images shown in [Figure 5] belong to the collection of the Arquivo Público Mineiro, Brazil.

9. REFERENCES


