Fusion of Local and Global Descriptors for Content-Based Image and Video Retrieval

Felipe S.P. Andrade, Jurandy Almeida,
Hélio Pedrini, and Ricardo da S. Torres*

Institute of Computing, University of Campinas – UNICAMP
13083-852, Campinas, SP – Brazil
felipe.andrade@students.ic.unicamp.br,
{jurandy.almeida,helio,rtorres}@ic.unicamp.br

Abstract. Recently, fusion of descriptors has become a trend for improving the performance in image and video retrieval tasks. Descriptors can be global or local, depending on how they analyze visual content. Most of existing works have focused on the fusion of a single type of descriptor. Different from all of them, this paper aims to analyze the impact of combining global and local descriptors. Here, we perform a comparative study of different types of descriptors and all of their possible combinations. Extensive experiments of a rigorous experimental design show that global and local descriptors complement each other, such that, when combined, they outperform other combinations or single descriptors.

Keywords: visual information retrieval, image and video descriptor, information fusion, genetic programming, performance evaluation.

1 Introduction

Recent advances in technology have increased the availability of image and video data, creating a strong requirement for efficient systems to manage those materials. Making efficient use of visual information requires the development of powerful tools to quickly find similar contents. For this, it is necessary (1) to design a feature extractor for encoding visual properties into feature vectors and (2) to define a similarity measure for comparing image and video data from their vectors, a pair known as descriptor [10].

In the literature, two main strategies have been considered for extracting visual features [7]: global descriptors, which are computed using the whole data and have the ability of generalizing the visual content with a single feature vector; or local descriptors, which are computed on boundaries between regions or points of interest.

To improve the performance of image and video retrieval systems, a new trend is the fusion of visual features. A common approach is the use of learning methods for combining different descriptors, such as Genetic Programming (GP) [6]. Most of existing works have focused on the fusion of a single type of descriptor (i.e., global or local) [5,9,12]. In spite of all the advances, it is still unclear how the

* Thanks to Brazilian agencies FAPESP, CNPq, and CAPES for funding.

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 845–853, 2012.

fusion of different types of descriptors affects the performance of those systems. A major difficulty of dealing with different types of descriptors is the different nature of data [6].

This paper aims to fill such a gap. Here, we perform a comparative study of different types of descriptors, including all of their possible combinations. Moreover, we carry out those analysis on the GP framework [11], which provides an effective way for combining descriptors. To the best of our knowledge, although this approach has been used for combining descriptors of the same type, it has never been employed in the fusion of different types of descriptors.

Extensive experiments were conducted on three large collections, comprising of both image and video data. Results from a rigorous experimental comparison of twelve descriptors, covering a variety of visual features, show that different types of descriptors act in a complementary manner, improving the performance in retrieval tasks.

The remainder of this paper is organized as follows. Section 2 briefly reviews the definition and the taxonomy of descriptors. Section 3 presents the GP framework and shows how to apply it for combining descriptors. Section 4 reports the results of our experiments and compares different combinations of descriptors. Finally, we offer our conclusions and directions for future work in Section 5.

2 Image and Video Descriptors

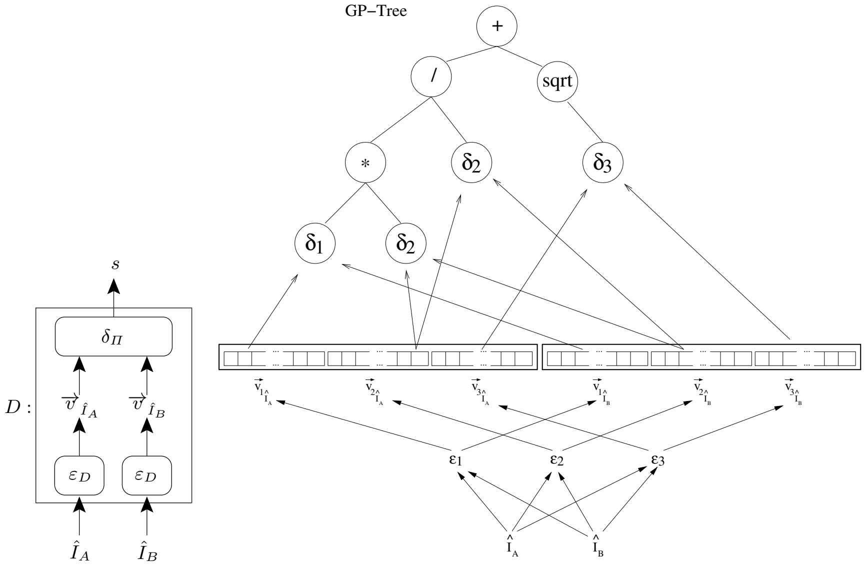
Both the effectiveness and the efficiency of image and video retrieval systems are very dependent on *descriptors*. A descriptor is responsible for characterizing visual properties and for computing their similarities, making it possible the ranking of images and videos based on their visual content.

Formally, a descriptor D can be defined as a pair (ϵ_D, δ_D) [10], where ϵ_D is a *feature-extraction algorithm* for encoding visual properties (e.g., color, shape, and texture) into feature vectors; and δ_D is a *similarity function* for comparing visual data from their corresponding feature vectors.

Figure 1(a) illustrates the use of a simple descriptor D to compute the similarity between two images (or videos) $\hat{I}_A$ and $\hat{I}_B$. First, the feature-extraction algorithm ϵ_D is used to compute the feature vectors $\overrightarrow{v}_{\hat{I}_A}$ and $\overrightarrow{v}_{\hat{I}_B}$ associated with $\hat{I}_A$ and $\hat{I}_B$, respectively. Next, the similarity function δ_D is used to determine the similarity value s between $\hat{I}_A$ and $\hat{I}_B$.

The feature-extraction algorithm ϵ_D can produce either a single feature vector or a set of feature vectors. In the former case, when a single feature vector must capture the entire information of the visual content, we say it is a *global descriptor*. In the latter case, a set of feature vectors is associated with different features of the visual content (regions, edges, or small patches around points of interest) and it is called *local descriptor*.

The similarity function δ_D can also be calculated as the inverse of a *distance function*. The most simple and widely used distance functions are the Manhattan distance (L_1) and the Euclidean distance (L_2). There are also more complex functions, such as the Earth Mover's Distance (EMD).



(a) Descriptor (b) GP-based method

Fig. 1. Comparison between a simple descriptor and a GP-based similarity function

3 The GP Framework

Genetic programming (GP) [6] is a set of artificial intelligence search algorithms designed following the principles of biological inheritance and evolution. The solution to a problem is represented as an individual in a population pool. The population of individuals evolves generation by generation through genetic transformation operations - such as reproduction, crossover, and mutation - with the aim at creating more diverse and better performing individuals with better fitness values in subsequent generations. A fitness function is available to assign the fitness value for each individual.

The entire discovery framework can be seen as an iterative process. Starting with a set of training data with known relevance judgements, GP first operates on a large population of random combination functions. These combination functions are then evaluated based on the relevance information from training data. If the stopping criteria is not met, it will go through the genetic transformation steps to create and evaluate the next generation population iteratively.

In this paper, we adopted the GP framework proposed by Torres et al. [11] for combining image and video descriptors to support queries based on visual content. In this context, such a framework can be used for building the so-called *Content-Based Image and Video Retrieval systems*. These systems can be characterized as follows: assume that we have a image (or video) database containing a large number of images (or videos). Given a user-defined query pattern (i.e., image or video), retrieve a list of the images (or videos) from the database which are most "similar" to the query pattern according to the visual content (i.e., the objects represented therein and their properties, such as shape, color, and texture, encoded through image and video descriptors).

Figure 1(b) illustrates the whole process to generate a GP-based similarity function. Let $\mathcal{D} = \{D_1 = (\epsilon_1, \delta_1), D_2 = (\epsilon_2, \delta_2), D_3 = (\epsilon_3, \delta_3)\}$ be a set composed by 3 pre-defined descriptors. First, their extraction algorithms ϵ_i are executed for each image (or video), and the resulting feature vectors are concatenated. In the following, a new similarity function is obtained by combining the similarity functions δ_i, through the GP framework. This new function can now be used to compute the similarity between images (or videos) $\hat{I}_A$ and $\hat{I}_B$, by using their feature vector representations. Usually, the computation of the GP-base similarity function is performed completely offline and, hence, does not impact on the search time. The overall framework is presented in Algorithm 1.

Algorithm 1. GP Framework

```
Generate an initial population of random "similarity trees"
while number of generations $\leq N_{gen}$ do
    Calculate the fitness of each similarity tree
    Record the top $N_{top}$ similarity trees
    for all the $N_{top}$ individuals do
        Create a new population by reproduction, crossover, and mutation operations.
    end for
end while
Apply the "best similarity tree" (i.e., the first tree of the last generation) on a set
of testing (query) data
```

The GP framework for the image and video retrieval problem is considered "global", in the sense it tries to find the best descriptor combination (represented as just one tree), which optimizes the number of the relevant results returned.

4 Experiments and Results

Experiments were conducted on three large collections: FreeFoto Nature, Caltech25, and Youtube10. FreeFoto Nature is a subset of the FreeFoto dataset[1] containing 3,461 natural images divided into 9 classes. Caltech25 contains 4,991 images from 25 classes of the Caltech101 dataset[2]. The number of images per class varies from 80 to 800. YouTube10 is a ten-class collection with 696 videos (245,402 frames) downloaded from YouTube[3] that we have created based on well-defined semantic categories. To reduce the amount of redundant information, a sampling method was used to extract one frame at each 2 seconds of video, totaling 115,082 frames. It is important to highlight that a better compression can be obtained by employing more elaborate techniques, such as summarization methods [2,3]. Examples of categories are Apple, Soccer, and World Trade Center. The average number of videos per class is 70.

[1] `http://www.freefoto.com/browse/205-00-0/Nature` As of April 2012.
[2] `http://www.vision.caltech.edu/Image_Datasets/Caltech101/` As of April 2012.
[3] `http://www.youtube.com` As of April 2012.

For describing those visual data, we used six global descriptors: ACC, BIC, GCH, and JAC, for encoding the color information; LAS and QCCH, for analyzing the texture property. For more details regarding those global descriptors, refer to [8]. In addition, we also extracted SIFT and SURF features, from which six local descriptors were generated. Two of them employ the EMD for comparing local features [4] and are referred to as SIFT and SURF, respectively. The others consist of a Bag-of-Words (BoW) representation defined by a 500-word vocabulary. When L_1 is used as distance function, they are named as SIFTBOF and SURFBOF; and for L_2, we call them L2SIFTBOF and L2SURFBOF. It is important to realize that any descriptor could be used for feature extraction.

To combine those descriptors, we evolved a population of 30 individuals along 10 generations using the GP framework (Section 3). For applying genetic operators, we used the tournament selection technique for selecting individuals. The reproduction, mutation, and crossover rates employed were 0.05, 0.2, and 0.8, respectively. A 5-fold cross validation was performed for each dataset in order to ensure statistically sound results. The reported results refer to the average scores obtained for the effectiveness measures.

We assess the effectiveness of each approach using Precision×Recall curves. Precision is the ratio of the number of relevant images/videos retrieved to the total number of irrelevant and relevant images/videos retrieved. Recall is the ratio of the number of relevant images/videos retrieved to the total number of relevant images/videos in the database. Some unique-value measurements are also used in the validation: Mean Average Precision (MAP), which is the area below Precision×Recall curves; and Precision at 5 (P@5), which is the average precision after 5 images/videos are returned.

Figure 2 presents the Precision×Recall curves observed for the different approaches. Those graphs compare the effectiveness of the global descriptors (first column) and the local descriptors (second column) for the FreeFoto (first row), Caltech (second row), and YouTube (third row) datasets. Our GP-based methods (third column) are also included: *GP-Global*, which combines the global descriptors; *GP-Local*, which combines the local descriptors; and *GP-GlobalLocal*, which combines both the global and local descriptors.

As we can observe, the GP-based similarity functions perform better than all the descriptors. Observe that *GP-Global* performs slightly better than the global descriptors. However, the same does not happen for the *GP-Local*, which did not significantly improve the effectiveness of the local descriptors. On the other hand, the combination of both the global and local descriptors is promising, yielding the best results. Note the superiority of *GP-GlobalLocal* when compared with the use of the best global and local descriptors in isolation.

Table 1 presents the comparison of descriptors and GP-based methods with respect to the MAP and P@5 measures. MAP is a good indication of the effectiveness considering all positions of obtained ranked lists. P@5, in turn, focuses on the effectiveness of the methods considering only the first positions of the ranked lists. For each dataset, we highlight the best result of each approach (i.e., global descriptors, local descriptors, and GP-based methods). Again, similar results to those observed for the Precision×Recall curves were obtained.

Paired t-tests were performed to verify the statistical significance of those results. For that, the confidence intervals for the differences between paired means of each class from the database were computed to compare every pair of approaches. If the confidence interval includes zero, the difference is not significant at that confidence level. If the confidence interval does not include zero, then the sign of the difference indicates which alternative is better.

Tables 2 and 3 present the confidence intervals (with a confidence of 95%) of the differences between the GP-based methods and the best global and local descriptors for the MAP and P@5 measures, respectively. In addition, we also include a comparison of *GP-GlobalLocal* with *GP-Global* and *GP-Local*.

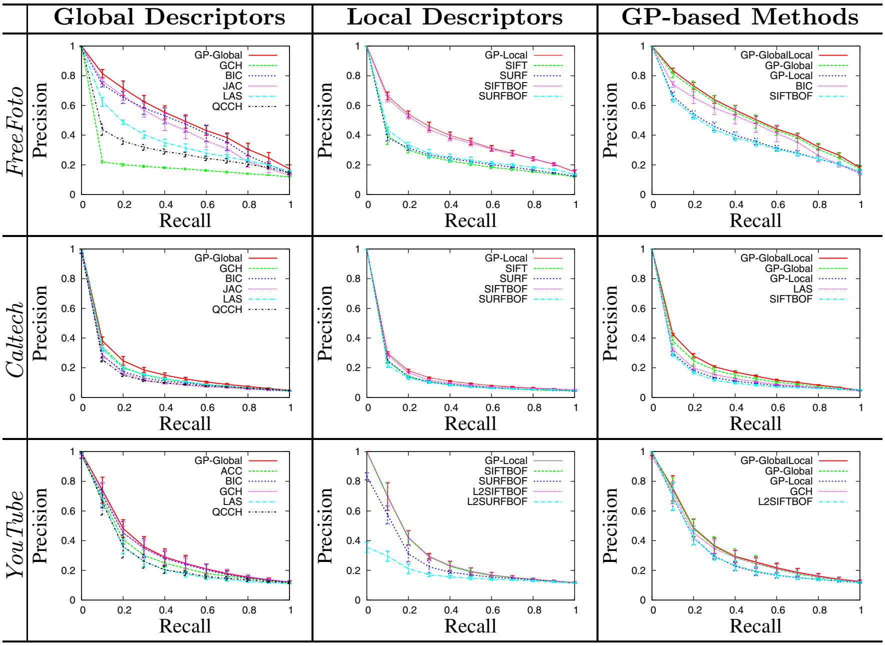


Fig. 2. Precision×Recall curves obtained by the global descriptors (first column), local descriptors (second column), and GP-based methods (third column) for the FreeFoto (first row), Caltech (second row), and YouTube (third row) datasets

The results show that *GP-Global* exhibits a similar performance to the best global descriptor, but is always better than the best local descriptor. As we also can see, the effectiveness of *GP-Local* is bounded by the poor performance of some local descriptors, reason for what the combination of them offers no performance gain. Despite the low effectiveness observed for the local descriptors, their combination with the global ones outperforms all the other approaches.

Figure 3 shows three query patterns and their top-5 results retrieved using the different approaches. The first position in each list of results is the query itself and the remaining ones are ranked in decreasing order of similarity regarding the query. The relevant results of each query are marked with a rectangular box.

Note that the global descriptors perform better than the local ones for the query *Starfish*, whereas the opposite behavior is performed for the query *Moun-*

Table 1. MAP and P@5 scores obtained by the global descriptors, local descriptors, and GP-based methods for the FreeFoto, Caltech, and Youtube datasets

	Approach	MAP			P@5		
		FreeFoto	*Caltech*	*YouTube*	*FreeFoto*	*Caltech*	*Youtube*
Global Descriptors	GCH	0.1596	0.1526	**0.3147**	0.1682	0.3249	**0.3934**
	BIC	**0.4648**	0.1681	0.3092	0.7668	**0.3599**	0.3902
	JAC	0.4453	0.1455	–	**0.7767**	0.3082	–
	ACC	–	–	0.2894	–	–	0.3575
	LAS	0.3575	**0.1703**	0.2635	0.6196	0.3500	0.3180
	QCCH	0.2914	0.1469	0.2673	0.4608	0.3076	0.3224
Local Descriptors	SIFT	0.2396	0.1350	–	0.4269	0.2920	–
	SURF	0.2477	0.1361	–	0.4247	0.2870	–
	SIFTBOF	**0.3802**	**0.1435**	0.2852	**0.6506**	**0.2950**	0.3534
	SURFBOF	0.2635	0.1253	0.2358	0.4760	0.2553	0.2965
	L2SIFTBOF	–	–	**0.2866**	–	–	**0.3544**
	L2SURFBOF	–	–	0.1547	–	–	0.1262
GP-based Methods	GP-Global	0.5063	0.1916	0.3215	0.8191	0.3838	0.4023
	GP-Local	0.3906	0.1582	0.2854	0.6656	0.3324	0.3523
	GP-GlobalLocal	**0.5211**	**0.2086**	**0.3280**	**0.8321**	**0.4051**	**0.4143**

Table 2. Differences between MAP of the different approaches at a confidence of 95%

Approach	*FreeFoto*		*Caltech*		*Youtube*	
	min.	max.	min.	max.	min.	max.
GP-Global – Best Global	-0.0201	0.1031	0.0091	0.0333	-0.0031	0.0168
GP-Global – Best Local	0.0617	0.1904	0.0049	0.0911	0.0067	0.0630
GP-Local – Best Global	-0.0503	0.0238	-0.0345	0.0103	-0.0603	0.0018
GP-Local – Best Local	0.0377	0.0273	-0.0102	0.0396	-0.0043	0.0019
GP-GlobalLocal – Best Global	-0.0229	0.1355	0.0131	0.0633	0.0052	0.0213
GP-GLobalLocal – Best Local	0.0887	0.1929	0.0139	0.1161	0.0148	0.0678
GP-GlobalLocal – GP-Global	-0.0078	0.0375	0.0024	0.0315	0.0001	0.0128
GP-GlobalLocal – GP-Local	0.0728	0.1881	0.0184	0.0822	0.0151	0.0699

Table 3. Differences between P@5 of the different approaches at a confidence of 95%

Approach	*FreeFoto*		*Caltech*		*Youtube*	
	min.	max.	min.	max.	min.	max.
GP-Global – Best Global	0.0139	0.0709	-0.0050	0.0529	-0.0040	0.0218
GP-Global – Best Local	0.0967	0.2402	0.0126	0.1650	0.0183	0.0777
GP-Local – Best Global	-0.1868	-0.0352	-0.0651	0.0100	-0.0721	-0.0101
GP-Local – Best Local	-0.0047	0.0348	-0.0214	0.0961	-0.0060	0.0060
GP-GlobalLocal – Best Global	0.0197	0.0910	0.0197	0.0707	0.0055	0.0362
GP-GLobalLocal – Best Local	0.1214	0.2415	0.0301	0.1900	0.0344	0.0855
GP-GlobalLocal – GP-Global	-0.0054	0.0313	0.0044	0.0380	-0.0002	0.0241
GP-GlobalLocal – GP-Local	0.1121	0.2207	0.0347	0.1107	0.0374	0.0865

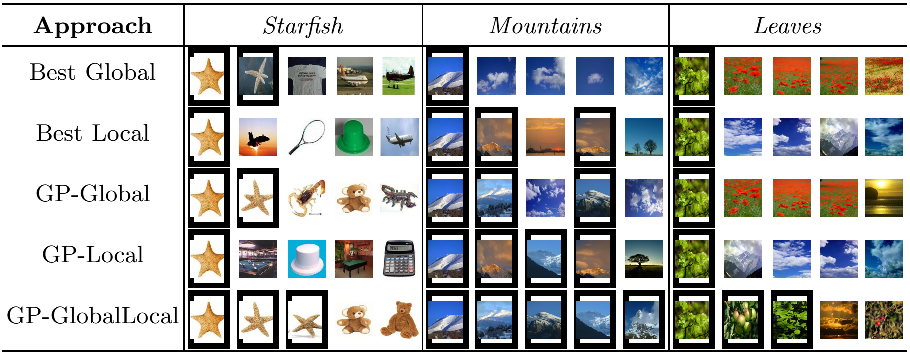


Fig. 3. Top-5 results retrieved by the different approaches in three queries: *Starfish* from Caltech, *Mountains* and *Leaves* from FreeFoto

tains. The results of both types of descriptors in isolation are improved by their combination, as shown by *GP-Global* and *GP-Local*. However, the same does not happen for the query *Leaves*, in which those approaches exhibit a poor performance. In spite of that, the fusion of the global and local descriptors takes the advantages of each type of descriptor, yielding significantly improved performance. Clearly, *GP-GlobalLocal* outperforms all the other approaches.

5 Conclusions

This paper has discussed the impact of combining different types of descriptors in the content-based image and video retrieval. Here, we have used a genetic-programming framework in the fusion of global and local descriptors.

We have conducted an extensive performance evaluation of twelve descriptors and all of their possible combinations, covering a variety of visual features. Results from a rigorous experimental comparison on three large datasets show that global and local features offer different and complementary information that can be exploited in order to improve the performance in retrieval tasks.

Future work includes the evaluation of other visual features for image and video retrieval (e.g., motion patterns [1]). We also plan to consider other learning-to-rank methods for combining global and local descriptors. Finally, we want to investigate the effects of their use in other applications.

References

1. Almeida, J., Leite, N.J., Torres, R.S.: Comparison of video sequences with histograms of motion patterns. In: Int. Conf. Image Proc. (ICIP), pp. 3673–3676 (2011)
2. Almeida, J., Leite, N.J., Torres, R.S.: VISON: VIdeo Summarization for ONline applications. Pattern Recognition Letters 33(4), 397–409 (2012)
3. Almeida, J., Torres, R.S., Leite, N.J.: Rapid video summarization on compressed video. In: Int. Symp. Multimedia (ISM), pp. 113–120 (2010)
4. Almeida, J., Rocha, A., Torres, R.S., Goldenstein, S.: Making colors worth more than a thousand words. In: Int. Symp. Appl. Comput. (SAC), pp. 1180–1186 (2008)

5. Ferreira, C.D., Santos, J.A., Torres, R.S., Gonçalves, M.A., Rezende, R.C., Fan, W.: Relevance feedback based on genetic programming for image retrieval. Pattern Recognition Letters 32(1), 27–37 (2011)
6. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press (1992)
7. Liu, Y., Zhang, D., Lu, G., Ma, W.Y.: A survey of content-based image retrieval with high-level semantics. Pattern Recognition 40(1), 262–282 (2007)
8. Penatti, O.A.B., Valle, E., Torres, R.S.: Comparative study of global color and texture descriptors for web image retrieval. J. Visual Commun. Image Representation 23(2), 359–380 (2012)
9. Salgian, A.: Combining local descriptors for 3D object recognition and categorization. In: Int. Conf. Pattern Recognition (ICPR), pp. 1–4 (2008)
10. Torres, R.S., Falcão, A.X.: Content-Based Image Retrieval: Theory and Applications. J. Theoretical and Applied Informatics 13(2), 161–185 (2006)
11. Torres, R.S., Falcão, A.X., Gonçalves, M.A., Papa, J.P., Zhang, B., Fan, W., Fox, E.A.: A genetic programming framework for content-based image retrieval. Pattern Recognition 42(2), 283–292 (2009)
12. Wu, Y.: Shape-based image retrieval using combining global and local shape features. In: Int. Congress Image and Signal Processing (CISP), pp. 1–5 (2009)