2010 Second International Conferences on Advances in Multimedia

Recognizing Landmarks Using Automated Classification Techniques: an Evaluation of Various Visual Features

Giuseppe Amato, Fabrizio Falchi, and Paolo Bolettieri
ISTI-CNR
via G. Moruzzi, 1 - Pisa, Italy
name.surname@isti.cnr.it

***Abstract*—In this paper, the performance of several visual features is evaluated in automatically recognizing landmarks (monuments, statues, buildings, etc.) in pictures. A number of landmarks were selected for the test. Pictures taken from a test set were classified automatically trying to guess which landmark they contained. We evaluated both global and local features. As expected, local features performed better given their capability of being less affected to visual variations and given that landmarks are mainly static objects that generally also maintain static local features. Between the local features, SIFT outperformed SURF and ColorSIFT.**

Keywords - Image indexing, image classification, recognition, landmarks.

I. Introduction

The amount of pictures taken by individuals has exploded during the last decade due to the wide adoption of the digital photography in the consumer market. However, many of these pictures remain unannotated and are stored with anonymous names on personal computers. Currently, there are no tools and effective technologies to help users in searching pictures by content, when they are not explicitly annotated. Therefore, it is becoming more and more difficult for users to retrieve even their own pictures.

A picture contains a lot of implicit conceptual information that, if understood how to be automatically inferred and used, can open up opportunities for new advanced applications. For instance, in addition to automatically create annotations and descriptions, pictures could also be used as queries on the web.

Given that smartphones equipped with cameras are becoming very popular nowadays, we can imagine that people, for instance tourists, can search for information on the web by simply pointing the camera of their smartphone on some subject (a monument, a restaurant, a painting). Consider in this respect the experimental service "Google Goggles" [1] recently launched by Google, that allows you to obtain information about a monument through your smarthphone using this paradigm.

Note that, even if many smartphones and cameras are equipped with a GPS and a compass, the geo-reference obtained with this is not enough to infer what the user is actually aiming at. Content analysis of the picture is still needed to determine more precisely the user query or the annotation to be associated with a picture.

In this respect, many researcher have been investigating the use of classification techniques as for instance, Support Vector Machines [2], k-Nearest Neighbor (k-NN) classifiers [3], boosting [4], etc., with visual information, with the purpose of automatically recognize visual content.

Content based retrieval and content based classification techniques typically are not directly applied to images content. Rather, matching and comparisons between low level mathematical descriptions of the images visual appearance, in terms of color histograms, textures, shapes, point of interests, etc., are used. Different visual features represent different visual aspects of an image. All together, different visual features, contribute, not exhaustively, to represent the complete information contained in an image. A single feature is generally able to carry out just a limited amount of this information. Therefore, its performance varies in dependence of the specific dataset used and the type of conceptual information one wants to recognize.

The goal of this paper is to identify the best visual features or combination of visual features that provides us with the best performance with the above mentioned task. In this respect, as better described in the reminder of the paper, we identified 12 landmarks, and we manually built the training sets for them by identifying a congruous number of pictures representing them.

A classification algorithm was tested with these landmarks, using various visual features. We measured the performance of the classification algorithm to correctly recognize the landmark in a test set, varying the visual features used.

The rest of the paper is organized as follows. We briefly discuss related work next. In Section III we present the features used in the experiments, while in Section IV we describe the experimental enviroment. Finally, we present and discuss the results in Section V.

II. Related work

In [5], the MPEG-7 Visual Descriptors have been compared in terms of effectiveness for a general purpose Content Based Image Retrieval (CBIR). The results are interesting

978-0-7695-4068-9/10 $26.00 © 2010 IEEE
DOI 10.1109/MMEDIA.2010.20

because real users were involved. However, the task proposed to the user was related to generic similarity search and not to recognition. Image classification based on MPEG-7 visual descriptors is addressed in [6]. The approach is very different from ours, since the authors choose to use a single learning algorithm which takes as input a single representation that combines the contributions of the individual MPEG-7 descriptors. In [7], various MPEG-7 descriptors have been used to build classifier committees. The focus of this paper is not on comparing the features but on using all of them at the same time. The committees have been tested on a slabs of stones dataset.

The first approach to recognizing location from mobile devices using image-based web search was presented in [8]. Two image matching metrics were used: energy spectrum and wavelet decompositions. Local features were not tested.

In [9], Google presented its approach to building a web-scale landmark recognition engine. Most of the work reported was used to implement the Google Goggles service. The approach makes use of the SIFT feature without comparing the performance of this feature with others.

An important survey of local features detectors is [10]. However, the various local features are not compared. In this paper we decided to use for each local feature the detector proposed by the authors of each feature.

III. Visual Features

In order to perform our evaluation we choose various global and local visual features. Specifically, we evaluated the performance of the 5 MPEG-7 [11] visual features (Color Layout, Color Structure, Edge Hystogram, Homogeneous Textures, Scalable Colour), the Scale invariant Feature Transform (SIFT) [12], the ColorSIFT [13], and the Speeded Up Robust Features (SURF) [14]. In the following we give a brief description of their principles.

A. MPEG-7

MPEG-7 visual descriptors consist of a set of 5 different global descriptors of the low level visual content of an image [11]. These 5 descriptors are mathematical representations of different statistical measures that can be computed analyzing the structure and placement of the colored pixel in an image. In particular:

- Scalable Color is an histogram of the colors of the pixel in an image, when colors are represented in the Hue Saturation Value (HSV) space
- Color Structure expresses local color structure in an image by use of a structuring element that is comprised of several image samples
- Color Layout is a compact description of the spatial distribution of colors in an image
- Edge Histogram descriptor describes edge distribution with a histogram based on local edge distribution in an image, using five types of edges
- Homogeneous Texture descriptor characterizes the properties of the texture in an image.

For extracting the MPEG-7 visual descriptors we made use of the MPEG-7 eXperimental Model (XM) Reference Software [15].

B. SIFT

The Scale Invariant Feature Transformation (SIFT) [12] is a representation of the low level image content that is based on a transformation of the image data into scale-invariant coordinates relative to local features. Local feature are low level descriptions of keypoints in an image. Keypoints are interest points in an image that are invariant to scale and orientation. Keypoints are selected by choosing the most stable points from a set of candidate location. Each keypoint in an image is associated with one or more orientations, based on local image gradients. Image matching is performed by comparing the description of the keypoints in images. For both detecting keypoints and extracting the SIFT features we used the public available software developed by David Lowe [16].

C. ColorSIFT

ColorSIFT local features [13] are an extension of the original SIFT definition to also take color into account. Basically, the original SIFT definition describes the local edge distribution around keypoints. The ColorSIFT extends the description of a keypoint also to colors around it. This is obtained by considering color gradients, rather than just intensity gradients. For both, detecting keypoints and extracting the Color SIFT features, we used the public available software developed by Jan-Mark Geusebroek [17]. Between the various proposals they made, we tested the colour-based SIFT invariant to shadow and shading effects which performed best in the experiments reported in [13].

D. SURF

The basic idea of Speeded Up Robust Features (SURF) [14] is quite similar to SIFT. SURF detects some keypoints in an image and describes these keypoints using orientation information. However, the SURF definition uses a new method for both detection of keypoints and their description that is much faster still guaranteeing a performance comparable or even better than SIFT. Specifically, keypoint detection relies on a technique based on a approximation of the Hessian Matrix. The descriptor of a keypoint is built considering the distortion of Haar-wavelet responses around the keypoint itself. For both, detecting keypoints and extracting the SURF features, we used the public available noncommercial software developed by the authors [18].

IV. Evaluation settings

A. The Dataset

The dataset is composed of 1227 photos of landmarks located in Pisa. The photos have been crawled from Flickr

Figure 1. Example images taken from the dataset

the well known on-line photo service. The dataset we built is publicly available. The IDs of the photos used for these experiments together with the assigned label and extracted features can be downloaded from [19]. The following is the list of labels assigned to the photos and the number of photos belonging to each class. In Figure 1 we reported an example for each class in the same order as they are reported in the below list:

- *Leaning Tower* (119 photos) – leaning campanile
- *Duomo* (130 photos) – the cathedral of St. Mary
- *Battistero* (104 photos) – the baptistery of St. John
- *Camposanto Monumentale (exterior)* (46 photos)
- *Camposanto Monumentale (field)* (113 photos)
- *Camposanto Monumentale (portico)* (138 photos)
- *Chiesa della Spina* (112 photos) – Gothic church
- *Palazzo della Carovana* (101 photos) – historic building
- *Palazzo dell'Orologio* (92 photos) – historic building
- *Guelph tower* (71 photos)
- *Basilica of San Piero* (48 photos) – church of St. Peter
- *Certosa* (53 photos) – the charterhouse

For the task of building and evaluating a classifier for the dataset classes, we divided the dataset in a training set consisting of 921 photos (approximately 80% of the dataset) and a test set consisting of 226 (approximately 20% of the dataset). The image resolution used for feature extraction is the standard resolution used by Flickr i.e., maximum between width and height equal to 500 pixels. However, for some specific experiments reported in the following we used lower resolution versions of the photos i.e., a maximum between width and height equal to 240 pixels.

B. Classification technique

Given a set of documents D and a predefined set of *classes* (also known as *labels*, or *categories*) $C = \{c_1, \ldots, c_m\}$, *single-label document classification* (SLC) [20] is the task of automatically approximating, or estimating, an unknown *target function* $\Phi : D \rightarrow C$, that describes how documents ought to be classified, by means of a function $\hat{\Phi} : D \rightarrow C$, called the *classifier*, such that $\hat{\Phi}$ is an approximation of Φ. In the experiments we present, a set of manually annotated documents has been partitioned into two subsets: the *training set* and the *test set*. The *training set* will be used in order to generate the classifiers $\hat{\Phi}$ by means of a supervised learning method, while the *test set* will be used in order to test the effectiveness (i.e., accuracy, precision, recall and F_1) of the generated classifiers.

The well-known *single-label distance-weighted k-NN* technique assigns a label in two steps. First it executes a k-NN search between the objects of the *training set*. The result of such operation is a list of labeled documents d_i belonging to the *training set* ordered with respect to the decreasing values of the similarity $s(d_x, d_i)$ between d_x and d_i. The label $\Phi_s(d_x)$ assigned to the document d_x by the classifier is the class $c_j \in C$ that maximizes the sum of the similarity between d_x and the documents d_i in the k-NN results list $\chi^k(d_x)$ labeled c_j. Formally the predicted class is:

$$\hat{\Phi}_s(d_x) = \arg\max_{c_j \in C} \sum_{d_i \in \chi^k(d_x) \ : \ \Phi(d_i) = c_j} s(d_x, d_i) \quad (1)$$

C. Similarity measures

For each feature used in the experiments we need a measure that evaluates the similarity between two photos. For the MPEG-7 visual descriptors we used the distances suggested by the MPEG Group in [15]. Let $d(d_x, d_y)$ be the distance, we defined the similarity between to objects as:

$$s(d_x, d_y) = 1 - w * d(d_x, d_y) \quad (2)$$

where w is a fixed number that guarantees that $w * d(x, y) < 1$ for any d_x and d_y.

In the experiments we also tested the weighted sum distance of these 5 MPEG-7 Visual Descriptors used in the *Search in Audiovisual using Peer-to-Peer Information Retrieval (SAPIR) FP6 European research project [21]*. More information about this combination can be found in [22].

The result of the comparison of two images d_x and d_y using local features (e.g., SIFT, ColorSIFT and SURF) is typically the number of keypoints in d_x that have a match in d_y. We translate this information in a similarity function dividing the number of matches by the number of keypoints in d_x. In other words we used the ratio of keypoints in d_x that do have a match in d_y as the similarity between d_x and d_y for all the local features used for the experiments (i.e., SIFT, ColorSIFT and SURF).

The algorithms used for matching the keypoints for the various local features are the ones suggested by the features authors and that are also used in their public available implementations. In particular both SIFT and ColorSIFT performs a 2-NN search between the keypoints in d_y for any keypoint in d_x. A match is identified if the 1st result in the 2-NN has a distance from the query keypoint less than 0.6 times the distance of the 2nd result. SURF matching algorithm is very similar except that the distance of the 1st nearest neighbor must be less than $1/\sqrt{2}$. More information can be found in [12], [13], [14].

D. Performance measures

For evaluating the effectiveness of the classifiers in classifying the documents of the $testset$ we use the micro-averaged *accuracy* and micro- and macro-averaged *precision*, *recall* and F_1.

Micro-averaged values are calculated by constructing a global contingency table and then calculating the measures using these sums. In contrast macro-averaged scores are calculated by first calculating each measure for each category and then taking the average of these. In most of the cases we reported the micro-averaged values for each measure. However, macro-averaged values for the best settings are reported in Figure 5.

Precision is defined as the ratio between correctly predicted and the overall predicted documents for a specific class. *Recall* is the ratio between correctly predicted and the overall actual documents for a specific class. F_1 is the harmonic mean of *precision* and *recall*.

Note that for the *single-label* classification task, micro-averaged *accuracy* is defined as the number of documents correctly classified divided by the total number of documents in the *test set* and it is equivalent to the micro-averaged *precision*, *recall* and F_1 scores.

V. Results

As explained in Section IV-B, the *single-label distance-weighted k-NN* technique has a parameter k. This parameter should be set during the training phase and kept fixed during the performance evaluation on the *test set*. However, in this paper we do not want to evaluate the specific technique but the relative performance of the various features using the same classification technique. In particular we do not want to test a specific training algorithm. Thus, we decided to report the performance measures we obtained for various k. In this way we can analyze the optimal performance (in the k range considered) and the stability changing the k parameter.

In Figure 2 we report the micro-averaged *accuracies* obtained for some MPEG-7 Visual Descriptors and their weighted sum combination used in the SAPIR Project (see IV-C). The best performance is obtained using the EdgeHistogram visual descriptor. The color-based features

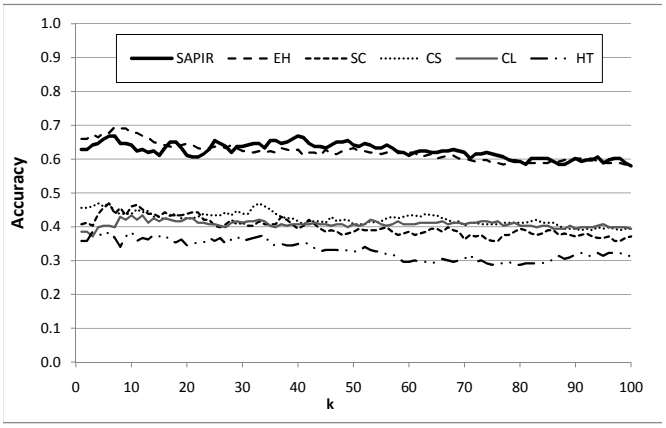

Figure 2. Micro-averaged accuracy of the classifier for various *k* and various global features (i.e., MPEG-7 Visual Descriptors and each weighted sum combination used in the SAPIR project)

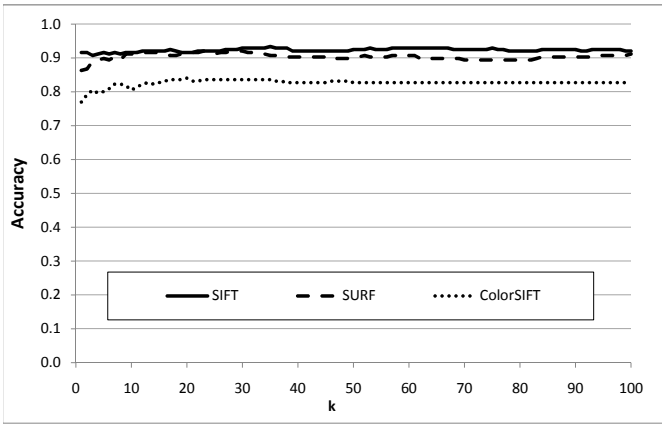

Figure 3. Micro-averaged accuracy of the classifier for various *k* and various local features

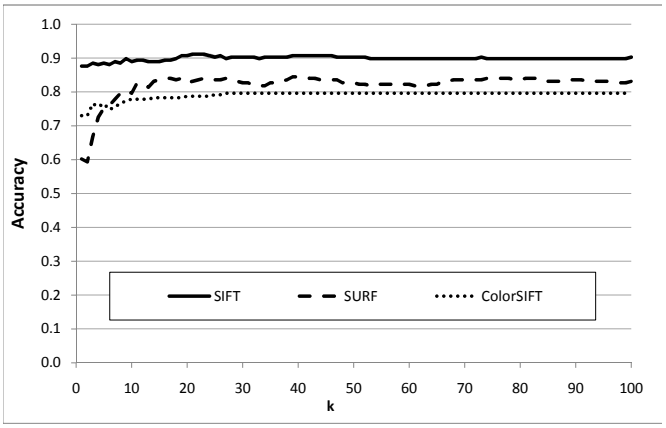

Figure 4. Micro-averaged accuracy of the classifier for various *k* and various local features. These experiments have been conducted using the low resolution version of the images

(i.e., ColorLayout, ColorStructure, ScalableColor) have very similar performance while HomogenousTexture obtained the worst values of *accuracy*. The weighted-sum combination of these visual descriptor performs slightly worst than EdgeHistogram alone. Even if for big values of k the SAPIR metric is preferable, the best *accuracy* for the various k is higher for EdgeHistogram alone.

The *accuracy* obtained for the local features are reported in Figure 3. As expected, all of them perform significantly better than the global features. In fact, the dataset used is specific for landmarks recognition and they are supposed to work well for general recognition tasks. What was not obvious is that SIFT (the oldest) perform better than the others. Both SURF and ColorSIFT are basically extensions of the SIFT but for this specific task they are less effective than SIFT. Even if in this paper we will not consider the computational cost, it is worth to mention that SURF is less computationally expensive than SIFT while ColorSIFT is even more demanding than SIFT.

Given the good results reported in Figure 3, we decided experimenting the same features on images with a lower resolution. We then extracted the same features from a lower resolution (i.e., max between width and height equal to 240 pixels) versions of the photos in the dataset. The results obtained, reported in Figure 4, show that the performance of SURF are more influenced by the resolution of the images even if their performance are still better then ColorSIFT.

Finally, we report in Figure 5 a complete summary of the performance obtained by each feature for the optimal k. The global features show a significant variance in the performance obtained for the various classes. Note that the macro- and micro-averaged results do not differ significantly and the previous considerations based on the micro-averaged *accuracy* would not change considering the macro F_1.

VI. Conclusions

In this paper we have performed a systematic evaluation of several visual features, considering as application the task of automatically recognizing known landmarks in a picture. The application was implemented as a classification task where every known landmark was identified by a class. We decide that a landmark is contained in a picture if the picture belongs to the class associated with the landmark.

The experiments that we carried out demonstrated the superiority of the local features over the global features. Specifically, the best results were obtained using the SIFT features that, even if proposed years before the other local features, is a bit better than SURF and ColorSIFT. However, we should mention that, even if we did not report any result concerning efficiency, algorithms for SURF are significantly faster than SIFT both when features have to be extracted and when they have to be matched.

Acknowledgment

This work was partially supported by the VISITO Tuscany project, funded by Regione Toscana, in the POR FESR 2007-2013 program, action line 1.1.d, and the MOTUS project, funded by the Industria 2015 program.

References

[1] "Google goggles," http://www.google.com/mobile/goggles/, last accessed on 30-March-2010.

[2] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, March 2000.

[3] T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13, no. 1, pp. 21–27, 1967.

[4] R. E. Schapire, "The boosting approach to machine learning an overview," in *Nonlinear Estimation and Classification*, Springer, Ed., 2003.

[5] G. Amato, F. Falchi, C. Gennaro, F. Rabitti, P. Savino, and P. Stanchev, "Improving image similarity search effectiveness in a multimedia content management system," in *Proc. of Workshop on Multimedia Information System (MIS)*, 2004, pp. 139–146.

[6] E. Spyrou, H. Le Borgne, T. Mailis, E. Cooke, Y. Avrithis, and N. O'Connor, "Fusing mpeg-7 visual descriptors for image classification," in *Proceedings of the 15th International Conference on Artificial Neural Networks (ICANN'05)*, 2005, pp. 847–852.

[7] T. Fagni, F. Falchi, and F. Sebastiani, "Adaptive committees of feature-specific classifiers for image classification," in *Image Mining. Theory and Applications. Proceedings of the 2nd International Workshop on Image Mining Theory and Applications (IMTA-09), Lisboa, Portugal, February 2009, INSTICC Press*, 2009, pp. 113–122.

[8] T. Yeh, K. Tollmar, and T. Darrell, "Searching the web with mobile images for location recognition," in *CVPR (2)*, 2004, pp. 76–81.

[9] Y. Zheng, M. Z. 0003, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven, "Tour the world: Building a web-scale landmark recognition engine," in *CVPR*. IEEE, 2009, pp. 1085–1092.

[10] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Found. Trends. Comput. Graph. Vis.*, vol. 3, no. 3, pp. 177–280, 2008.

[11] ISO/IEC, "Information technology - Multimedia content description interfaces," 2003, 15938.

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

[13] G. J. Burghouts and J. M. Geusebroek, "Performance evaluation of local colour invariants," *Computer Vision and Image Understanding*, vol. 113, pp. 48–62, 2009.

		Labels																				macro avg	micro avg
		Torre		Duomo		Battistero		CM-Ext		CM-Field		CS-Port.		Spina		Carovana		Orologio		Guelph T.		San Piero		Certosa			
MPEG-7 CL (k=10)	Precision	0.48		0.27		0.35		0.17		0.67		0.74		0.31		0.33		0.19		0.83		0.00		0.25		0.38	0.43
	Recall	0.50		0.35		0.29		0.11		0.86		0.87		0.41		0.35		0.22		0.36		0.00		0.09		0.36	
	F_1	0.49		0.31		0.32		0.13		0.75		0.80		0.35		0.34		0.21		0.50		0.00		0.13		0.35	
MPEG-7 CS (k=4)	Precision	0.38		0.42		0.26		0.60		0.76		0.82		0.29		0.44		0.36		0.41		0.57		0.33		0.47	0.47
	Recall	0.33		0.42		0.24		0.33		0.68		0.78		0.27		0.70		0.50		0.50		0.40		0.18		0.45	
	F_1	0.36		0.42		0.25		0.43		0.72		0.80		0.28		0.54		0.42		0.45		0.47		0.24		0.45	
MPEG-7 EH (k=7)	Precision	0.77		0.55		0.52		0.33		0.92		0.95		0.71		0.65		0.77		0.43		0.75		0.67		0.66	0.69
	Recall	0.83		0.62		0.67		0.11		0.82		0.78		0.77		0.75		0.94		0.43		0.60		0.36		0.63	
	F_1	0.80		0.58		0.58		0.17		0.87		0.86		0.74		0.70		0.85		0.43		0.67		0.47		0.64	
MPEG-7 HT (k=3)	Precision	0.33		0.24		0.28		0.29		0.78		0.53		0.20		0.52		0.56		0.25		0.38		0.09		0.37	0.38
	Recall	0.42		0.23		0.24		0.22		0.64		0.43		0.27		0.70		0.50		0.21		0.30		0.09		0.35	
	F_1	0.37		0.24		0.26		0.25		0.71		0.48		0.23		0.60		0.53		0.23		0.33		0.09		0.36	
MPEG-7 SC (k=6)	Precision	0.43		0.37		0.30		0.08		0.74		0.55		0.38		0.52		0.53		0.58		0.67		1.00		0.52	0.47
	Recall	0.42		0.50		0.29		0.11		0.61		0.74		0.36		0.70		0.44		0.50		0.40		0.09		0.43	
	F_1	0.43		0.43		0.29		0.10		0.67		0.63		0.37		0.60		0.48		0.54		0.50		0.17		0.43	
MPEG-7 SAPIR (k=6)	Precision	0.66		0.64		0.56		1.00		0.64		0.96		0.52		0.79		0.52		0.69		1.00		0.00		0.67	0.67
	Recall	0.88		0.62		0.43		0.11		0.89		1.00		0.50		0.75		0.83		0.79		0.40		0.00		0.59	
	F_1	0.75		0.63		0.49		0.20		0.75		0.98		0.51		0.77		0.64		0.73		0.57		0.00		0.58	
SIFT (k=35)	Precision	1.00		1.00		1.00		1.00		0.93		1.00		0.96		1.00		0.69		0.73		1.00		1.00		0.94	0.93
	Recall	1.00		0.96		1.00		0.67		1.00		0.74		1.00		1.00		1.00		0.79		0.90		0.91		0.91	
	F_1	1.00		0.98		1.00		0.80		0.97		0.85		0.98		1.00		0.82		0.76		0.95		0.95		0.92	
SIFT LR (k=21)	Precision	0.96		0.96		0.87		0.86		0.93		0.94		1.00		0.80		0.82		0.85		1.00		1.00		0.91	0.91
	Recall	0.96		0.92		0.95		0.67		1.00		0.70		0.95		1.00		1.00		0.79		0.90		0.91		0.89	
	F_1	0.96		0.94		0.91		0.75		0.97		0.80		0.98		0.89		0.90		0.81		0.95		0.95		0.90	
ColorSIFT (k=20)	Precision	0.80		0.96		0.77		1.00		0.80		0.94		0.95		1.00		0.76		0.60		0.88		0.78		0.86	0.84
	Recall	0.83		0.96		0.81		0.56		1.00		0.70		0.82		0.95		0.89		0.86		0.70		0.64		0.81	
	F_1	0.82		0.96		0.79		0.71		0.89		0.80		0.88		0.97		0.82		0.71		0.78		0.70		0.82	
ColorSIFT LR (k=27)	Precision	0.86		0.96		0.65		0.56		0.90		0.84		0.66		0.79		0.77		0.77		1.00		0.86		0.80	0.80
	Recall	0.79		0.88		0.71		0.56		0.96		0.70		0.86		0.95		0.94		0.71		0.40		0.55		0.75	
	F_1	0.83		0.92		0.68		0.56		0.93		0.76		0.75		0.86		0.85		0.74		0.57		0.67		0.76	
SURF (k=23)	Precision	0.80		1.00		1.00		1.00		0.97		0.71		0.92		1.00		1.00		1.00		1.00		0.90		0.95	0.92
	Recall	1.00		0.96		1.00		0.67		1.00		0.87		1.00		0.95		0.94		0.79		0.60		0.82		0.88	
	F_1	0.89		0.98		1.00		0.80		0.98		0.78		0.96		0.97		0.97		0.88		0.75		0.86		0.90	
SURF LR (k=39)	Precision	0.88		0.71		1.00		0.83		0.68		0.86		0.95		0.86		1.00		1.00		0.86		1.00		0.89	0.85
	Recall	0.88		0.96		0.95		0.56		0.96		0.83		0.86		0.95		1.00		0.43		0.60		0.55		0.79	
	F_1	0.88		0.82		0.98		0.67		0.79		0.84		0.90		0.90		1.00		0.60		0.71		0.71		0.81	

Figure 5. Precision, recall and F_1 for each class label plus macro- and micro-averaged. Note that micro-averaged *precision*, *recall* and F_1 are equals and equivalent to the micro-averaged accuracy. LR indicates the features extracted from the lower resolution version of the images

[14] H. Bay, T. Tuytelaars, and L. J. V. Gool, "Surf: Speeded up robust features," in *ECCV (1)*, 2006, pp. 404–417.

[15] ISO/IEC, "Information technology - Multimedia content description interfaces. Part 6: Reference Software," 2003, 15938-6:2003.

[16] "SIFT keypoint detector," http://people.cs.ubc.ca/∼lowe/, last accessed on 30-March-2010.

[17] "Color SIFT detector," http://staff.science.uva.nl/∼mark/, last accessed on 30-March-2010.

[18] "SURF detector," http://www.vision.ee.ethz.ch/∼surf/, last accessed on 30-March-2010.

[19] "Pisa landmarks dataset," http://www.fabriziofalchi.it/pisaDataset/, last accessed on 30-March-2010.

[20] S. Dudani, "The distance-weighted k-nearest-neighbour rule," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-6(4), pp. 325–327, 1975.

[21] "Search in Audiovisual using Peer-to-Peer Information Retrieval (SAPIR) FP6 European research project," http://www.sapir.eu.

[22] M. Batko, F. Falchi, D. Novak, R. Perego, F. Rabitti, J. Sedmidubsky, and P. Zezula, "Building a web-scale image similarity search system," *Multimedia Tools and Applications*, vol. to appear.