Computer Vision Winter Workshop 2006, Ondřej Chum, Vojtěch Franc (eds.)
Telč, Czech Republic, February 6–8
Czech Pattern Recognition Society

Extraction of Attributes, Nature and Context of Images

Stefan Kuthan and Allan Hanbury

Pattern Recognition and Image Processing Group
Institute of Computer Aided Automation, Vienna University of Technology
Favoritenstrasse 9/183-2 A-1040 Vienna, Austria
stefan.kuthan@gmx.at, hanbury@prip.tuwien.ac.at

Abstract *In this paper a framework for deriving high-level scene attributes from low-level image features is presented. Examples of attributes derived are photo-painting, indoor-outdoor, night-day and nature-city. The assignment of the attributes to images is done by a hierarchical classification of the low level features, which capture colour, texture and spatial information. A prototype for image classification is implemented, which aids in the evaluation of the different methods available. We give a detailed analysis of the indoor-outdoor cassification.*

1 Introduction

The automatic derivation of semantically-meaningful information from the content of an image is the focus of interest for much research on image databases.

This paper concerns the extraction of image semantic types (e.g. landscape photograph, clip art) from low-level image features. The extracted semantics could, for example, be used in conjunction with an automatic segmentation of images to guide the segmentation algorithm. The image database used for training and testing consists of 5474 images provided by the ImagEVAL project[1], a French computer vision evaluation project. Image semantics to be extracted, as specified in the project description, are:

1. Black and White - Colour - Manually Coloured
2. Photograph - Art/Painting
3. Outdoor - Indoor
4. Night - Day
5. City - Nature/Countryside

Figure 1 shows an example of the image annotation that is the goal in this paper.

A variety of applications for image classification and feature extraction can be found in *Content Based Image Retrieval* (CBIR). An application especially suited to the classification under consideration here is the automatic colour correction of consumer photos during film development [5][7]. Another application could be the automatic classification of images in large electronic-form art collections, such as those maintained by museums or image archives of print media / television. Generally speaking, such a classification is useful everywhere where a manual classification or sorting process is infeasible because of the number of images under consideration.

There exists much work on this sort of image classification [1][3][7][8][9][11], however papers often concentrate on a small subset of the classes given or even just a binary classification. Each evaluation is usually done on a different set of images, making it difficult to judge the effectiveness of the methods. This paper contributes by analysing the effectiveness of a large number of features for the tasks listed above. An effective feature combination method and hierarchical clustering approach is presented.

Section 2 presents an overview of the features extracted, while section 3 describes the classification methods used. The results are presented and discussed in section 4.

2 Features

All input images are encoded in the RGB colour space. Therefore it would be of advantage to work with RGB since no conversion is needed. The drawback however is that this space is ill-suited for most classification based on colour. For example, different illumination will change the perceived colour. While the human eye will make adjustments to accommodate for this, it is hard to construct a metric for which an image has the same (pixel) values regardless of lighting conditions. The luminance information is more important to our perception than the chroma, a difficult fact to consider when using a colour-space where luminance is not directly available, rather being a combination of all three channels.

To capture colour information, histograms are calculated in several colour spaces. This section shows why the particular conversions were considered and details on the parameters chosen. The number of bins per channel is 20.

RGB Histogram Although the RGB space was expected to perform worse than other colour spaces for the reasons mentioned above, there are good reasons for calculating a feature vector based on this space. An advantage is that no conversion errors are introduced. The classification of images into the nature and urban class was also expected to benefit from this space when considering the green channel

[1] project description: http://www.imageval.org

(a) colour, outdoor, day, nature (b) colour, outdoor, night, urban

Figure 1: Examples of semantic annotation

which is expected to show higher values for the nature class.

Ohta Histogram The Ohta colour space is proposed for indoor-outdoor classification in [7]. The first channel of this space captures brightness information as it is the sum of the three channels of RGB.

CIELUV/CIELAB Histogram An advantage of the CIELUV as well as the CIELAB colour space is that the Euclidean distance between two colours models the human perception of colour difference. The luminance information is directly available in the first channel.

srgb Histogram The calculation of the normalized RGB colour space[2] is performed as proposed in [1]. The "intensity free" image is computed by dividing each channel of RGB by the intensity at each pixel. The calculation of the intensities is as follows:

$$I = (299 * R + 587 * G + 114 * B)/1000 \qquad (1)$$

HSV The HSV colour space, representing hue, saturation and colour value (brightness) has the shape of a hexagonal cone. The angle is given by the hue, the distance from the centre of the cone by the saturation and the vertical position by the value. This colour space is used for a part of the **colour statistics** shown in the following list:

- Illuminant: this value indicates the colour of the light source. It is calculated in two versions, through the "Grey-world algorithm" and the "White patch algorithm". The former is calculated by the mean of the three colour channels, which is assumed to be "grey" (multiplied by 2 to get white), the latter is calculated by assuming that a white patch is always visible in an image, therefore taking the maximum value of each channel.

[2]This is not the sRGB as defined by IEC 61966-2-1 "Default RGB Colour Space".

- Unique colours: this value is calculated by transformation into the HSV-space and counting the unique values in the Hue channel.
- Histogram sparseness: a histogram is calculated and bins containing counts higher than a fixed cut-off value counted.
- Pixel saturation: this is calculated as a ratio between the number of highly saturated and unsaturated pixels in the HSV colour space [1].
- Variance in and between each channel of the RGB space.

The following texture features are calculated:

Edge direction This feature is used to compare the frequency of occurrence of edge directions. As with colour, a histogram is used to discretise the values. For a greyscale image the gradient is calculated in two directions by convolution with the horizontal and vertical Prewitt kernels. The next step is the calculation of the magnitude and direction at each pixel x:

$$m(x) = \sqrt{f_h(x)^2 + f_v(x)^2} \qquad (2)$$

$$\theta(x) = arctan\left(\frac{f_v(x)}{f_h(x)}\right) \qquad (3)$$

where f_h and f_v are the horizontal and vertical edges.

Edge direction coherence vector The calculation of the edge direction coherence vector is accomplished by a morphological closing of the magnitude image with a line segment followed by a morphological opening with a small disk. Thereby the dominating structures are enforced while degenerate "edges" – isolated pixels – are removed. As above, a greyscale image is used for the input. In both cases the result is a histogram of the direction image multiplied (masked) by the thresholded magnitude image. The 37 bins represent 5 degree intervals from -90 to 90 degrees. The number of edge pixels found is stored in an extra bin of the histogram. Normalization with the image size is also performed.

Edge Statistics This feature is used to determine whether the edges in the image result from intensity changes, as is the case with natural images, or from changes in hue, a method employed in paintings [1]. The intensity edges are found as above. The colour edges are found by first transforming the image into the srgb space, resulting in normalised RGB components. The colour edges of the resulting "intensity-free" image are then determined by applying the edge detector to the three colour channels and fusing the results by taking the maximum. The feature extracted is the fraction of pure intensity-edge pixels.

Wavelets The Haar transform[5] is used to decompose an image into frequency bands. To extract an image feature this transform is applied to the L component of a LUV image. The square root of the second order moment of wavelet coefficients in the three high-frequency bands is computed. This image feature captures variations in different directions. In the implementation of the prototype 4 levels are computed. This yields a feature vector of length 12.

Gabor filter The Gabor filter is a quadrature filter. It selects a certain wavelength range (bandwidth) around the centre wavelength using the Gaussian function. This is similar to using the windowed Fourier transform with a Gaussian window function. The feature vector is constructed by calculating the mean and standard deviation of the magnitude of the transform coefficients at several scales and orientations [6] [10]. This means that the fast Fourier transform (FFT) is applied to an image and then the Gabor filter, specific to this scale and orientation, is applied. Now the inverse of the FFT is taken and the mean and standard deviation calculated. For the prototype this filter is applied at 6 orientations and at 4 scales. Two values are collected at each point; therefore the feature vector has the length 48.

3 Classification

For implementation of the prototype Matlab Version 6.5 was used. The library PRTools[3] Version 4.0.14 [2] is used to construct the classifier. The results reported in the next section were obtained with the *k*-NN classifier, where the number of neighbours is set to 5. Other tested classifiers are deselected due to their complexity, sharply increasing computation time (neural net, Mixture of Gaussians), or because of their lower performance, probably because of the inability to model complex distributions (Linear and Quadratic Bayes and Parzen classifier). The Bagging classifier, based on *k*-NN and the Decision trees proved to be competitive but not as robust as the *k*-NN classifier.

3.1 Spatial Information

To capture spatial information, each image is divided into 16 sub-images. This 4×4 image tessellation is of benefit because image regions can be weighted according to their importance. For each sub-block a feature vector is calculated separately. A simple concatenation of these would increase the dimensionality by a factor of 16. To keep the classification simpler the following method is used: a classifier is built for each sub-block and a combining classifier, described in the next section, effectively weights the results of these.

A drawback of this approach is that only simple concepts can be captured through this method (e.g. blue sky at the top - for outdoor images). Complex concepts, such as XOR cannot be solved. As an example for successful weighting, Figure 2 shows the error rate for indoor-outdoor classification based on the RGB histogram, averaged over the sub-blocks of 1000 test images when trained with 2000 images. In Figure 2, white represents the best error rate of 0.244% and black the worst with 0.365%. As can be observed the classification is better for the blocks in the upper part of the images, probably capturing the "sky" information. Also the combination of the results of the individual sub-blocks brings an improvement to an overall error rate of 0.183%.

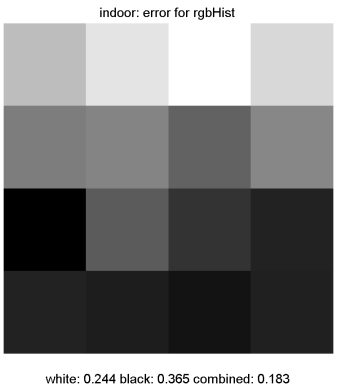


Figure 2: Using image tessellation to capture Spatial Information: Indoor-Outdoor

3.2 Combining Features

The method used for incorporating spatial information is extended for several features straightforwardly. For each sub-block and for each feature a classifier is trained using a subset of 70% of the data available. Depending on the number of features used, between 16 (for one feature) and 64 (for 4 features) classifiers have to be trained.

The training of the sub-blocks with 70% of the data is done to introduce "unseen" data for the combining classifier. This avoids overfitting the combining classifier.

The output when applying a classifier is a value signifying the confidence with which each image belongs to the class under consideration. The trained classifiers are applied to all of the training data independently.

In the next step their outputs are concatenated to a feature vector and the combining classifier trained. The number of classifiers for each sub-problem is therefore the number of blocks times the number of features plus one.

Experiments were also carried out with the possibilities for combining classifiers provided by PRTools. These are: Product, Mean, Median, Maximum, Minimum and Voting combiner. However classification with these combiners generally shows an error rate higher than that achieved with the scheme above.

[3] Pattern Recognition Tools: available at http://www.prtools.org/

3.3 Hierarchical Classification

A hierarchical classification similar to that described in [8] is implemented. The classifier for the whole problem is organised in the hierarchy shown in Figure 3.

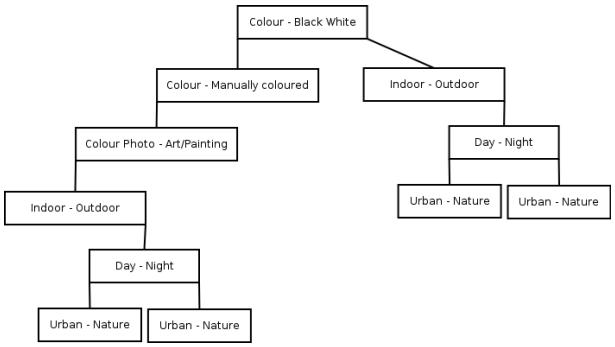


Figure 3: Hierarchy of Classifiers

At each node the training or application of a classifier takes place. Only the appropriate sub-sample of images, as determined by the node, is passed to the children nodes. At leaf nodes training or classification stops. This is a divide and-conquer strategy with several advantages. One advantage, compared to a classification of all attributes at once, is reduced complexity through reduction to two-class problems. Also there is no need for a third class of images belonging to none of the classes under consideration.

Each node can be configured individually. The prototype currently has settings for: enabling/ disabling classification, list of low-level features selected, prior probabilities, chosen combining scheme (classifier, voting scheme) and the list of children, if any. This structure could be extended for parameters specifying the type of classifier (*k*-NN, decision trees etc.) and parameters to use. During the training phase the obtained classifiers are also stored in this structure.

This scheme also helps to keep the feature-vector used for training and during classification as small as possible, for example for day-night classification only one feature is used.

The logic of the problem-domain is easy to implement through the setting of the "children" list. This allows for a relatively easy extension to other attributes. Through this integration of the logic, inherent in the targets, a plausibility-check is not needed for the class labels (e.g. a setting of two contradicting labels does not lead to an error). The hierarchy shown in Figure 3 was obtained through analysis of the problem domain.

When applying the classifier, classification stops at the leaf nodes. This leads to an increase of speed and could be further exploited to only extract the needed features for each image.

Each of the nodes can be analysed separately. Figures such as the one shown in Figure 2 are available for each attribute and feature pair and help to interpret performance at each node.

4 Results

For the comparison of features and also as a means to test their variance, box plots were created with a (smaller) sample of 700 training and 200 test images. Figure 4 shows an example of such a box plot for the outdoor-indoor classification. These results are used to manually select the best features for each sub-problem. Table 2 shows the medians of the hit-rates achieved with the features under consideration for each of the classes.

For the evaluation of the prototype a sample size of 2000 images is chosen for training and 1000 images are used for testing. The sample sizes were chosen for the purpose of faster testing, similar results are obtained when testing on the remaining 2474 images.

4.1 Overall Results

Table 1 shows the obtained hit-rate on classification. The features were chosen by analysis of the box plots. The baseline is calculated by division of the size of the bigger class by the total number of instances. This is the best result possible when guessing the class, without any feature available.

problem	features	hit-rate	baseline
BW	Lab colStat	99.0%	79.7%
Man.col.	Lab	96.1%	93.4%
Art	srgb wav	94.9%	91.4%
Photo	all of above	93.3%	64.6%
Indoor	rgb Lab edgeC wav	83.5%	63.5%
Night	Luv	96.5%	86.8%
Nature	rgb edgeC wav	87.1%	63.2%
Total		**71.0%**	20.0%

Table 1: Summary of Results

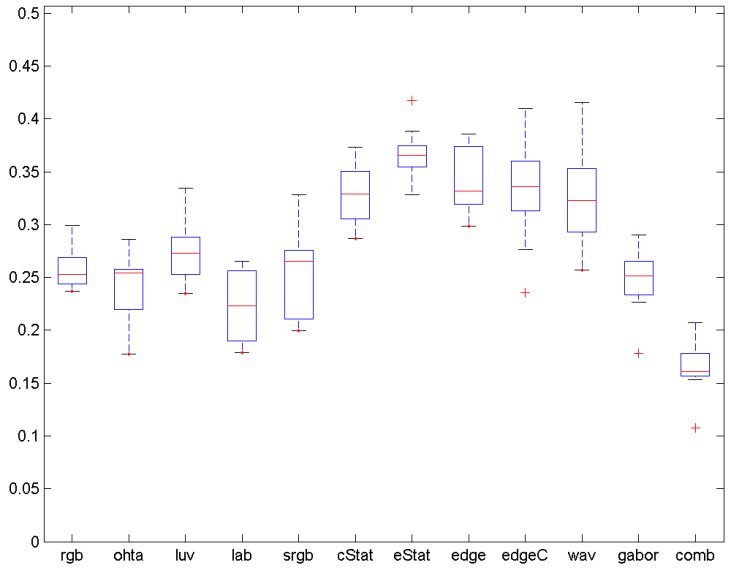
Some comments are given below, while a detailed analysis of one of the more interesting tasks, indoor-outdoor classification, is given in the next section[4].

Black and white images have little variance between the channels, a small error is made though the use of sepia images. Manually coloured images have a colour distribution that is uncommon in natural images. The classifier for the attribute art draws on the observation that natural images have more structure and a more even colour distribution than images of this class. Images that are not specified as belonging to any of the classes mentioned this far, are classified as colour photo. The error for assignment of the attribute night can be traced to ambiguous images, taken at dusk, with dark sky or underwater. Classification of images into nature and urban classes is based on strong vertical and horizontal structures in urban scenes and on colour differences, generally nature images have higher values in all three channels of the RGB histograms. The total percentage of correctly classified images (all attributes correctly assigned) is 71.0%.

4.2 Outdoor - Indoor

As an example of how the results obtained with the prototype are interpreted to select the best features, this section gives detail on the outdoor-indoor classification. A compact

[4]Detailed discussion of the results on all attributes are available in [4].



On the *x-axis* following features are shown: histograms in five colour spaces, colour statistics, edge statistics, edge direction histogram and coherence vector, wavelets, Gabor filter and the combination of all features.
On the *y-axis* the error-rate can be read off. Each box is limited by the lower quartile (25% of the data) and the upper quartile (75%). The median is indicated by a horizontal line. Whiskers and crosses show the extent of remaining data.

Figure 4: Example box plot: indoor-outdoor error for available features

Attribute	rgb	ohta	luv	lab	srgb	colour-stats
art	91.9	92.9	91.5	91.2	***93.1**	91.3
blackWhite	87.4	97.6	95.1	98.0	93.5	***98.6**
bw_Col	93.6	95.1	94.7	***95.7**	94.7	92.5
day	95.9	93.6	***96.2**	91.6	91.8	91.1
indoor	74.2	75.7	72.4	***77.7**	74.8	67.1
nature	*80.9	78.4	78.6	77.8	71.3	61.9

Attribute	edge-stats	edges	edges C.	wavelet	gabor	combined
art	91.6	*91.8	*91.8	91.2	*91.8	93.3
blackWhite	*89.3	76.1	74.4	73.7	76.9	98.1
bw_Col	93.0	93.0	92.9	92.5	*93.2	95.6
day	85.9	87.3	86.1	*88.6	85.9	96.8
indoor	63.4	65.8	66.9	67.6	*75.4	83.5
nature	57.6	78.8	77.0	75.4	***84.2**	88.1

Table 2: Comparison of Features - the percentage of correctly classified images is given; top: colour histograms, bottom: texture features. The best single feature is bold, an asterisk marks best result of each sub-table.

way of observing the errors on each class is the Confusion Matrix:

outdoor	indoor	<– classified as
456	69	outdoor
68	237	indoor

For the classification of images into the indoor or outdoor class the following features are selected: RGB and CIELAB histograms, coherent edge direction histogram and the wavelet filters. The result obtained in the classification process (83.5%) is not as good as that obtained on the nature of the image (bw, coloured bw, art or colour photo). However the results obtained by other authors (82% to 93%) are comparable because their training and test sets are often smaller and ambiguous images are eliminated beforehand. An interpretation of the box plot (Figure 4) is that generally colour features seem to perform better than texture features, what is striking is that the combination of all features yields a much better result than any single feature. Also the Gabor filter performs as well as the colour features.

An analysis of the RGB and CIELAB histograms shows that indoor images have slightly less luminance and (therefore) less highly saturated pixel values. Also the sub-block classifiers for these features perform better for the upper half of the images (Figure 2), this can be attributed to the presence or absence of a sky or alternatively that this area best reflects lighting conditions. The values obtained through the Gabor filters show higher values for the indoor class, indicating more structure or highly textured images. For the final implementation of the prototype the Gabor filter was deselected because of its high computational costs, however the wavelet filters, selected instead, show a similar response for this classification. As with the Gabor filter the result of the wavelet operation shows higher values for the indoor class. The coherent edge direction histograms show higher values for the outdoor class, seemingly contradicting this observation. Both classes show peaks at the values indicating horizontal, vertical and diagonal structures -90, -45,

0, 45 and 90 degrees. This effect is somewhat more pronounced for the indoor class.

The results obtained in combining the said features are similar to the combination of all features, as indicated by the feature "comb" in the box plot. It has been indicated in several papers that a combination of features has most effect when combining features of the "colour" group with those of the "texture" group.

The reason for indoor images to be classified as outdoor often seems to be lighting conditions caused by the presence of windows or doors. Also a strong presence of green or, in the case of black and white images, a bright background seems to bias the images into this class. The outdoor images classified as indoor either show very high detail or cluttering of the image or depict outdoor scenes with lighting common to indoor images, e.g. during dawn and dusk.

5 Conclusion

Reasonable results can be obtained in extracting image semantics with the aid of statistical methods. An accuracy of 71% is achieved on the problem posed by the ImagEVAL project. The hierarchical classification makes use of knowledge about the problem-domain. The attributes to be assigned to the images are mutually exclusive and cover a wide spectrum of input images.

The prototype developed is used for two aims. The result of image classification is used to evaluate and compare the discrimination power of several features on the given problems. Secondly, conclusions about the reasons why particular features are suited for a problem are drawn. This is done through an analysis of results and variables available at substages during the training and testing phases.

The ambiguity of natural language, where, for example, "nature" and "urban" is not explicitly defined and leads to problems in classification of images that cannot be accurately described with either word, is an unsolved problem.

To make use of the biggest image collection in the world, the Internet, the implementation of more attributes would be of interest. For example business graphic - photograph. The integration of meta-information of the images, e.g. size of file or information stored in Exchangeable Image File Format (EXIF) tags, would also be of interest.

An improvement of feature extraction speed would be of advantage, not only in use of the prototype with large image databases, but also to rapidly test other parameter settings and low-level features. The results obtained with the prototype are comparable to those found in literature.

Acknowledgement

This work was supported by the European Union Network of Excellence MUSCLE (FP6-507752), and the Austrian Science Foundation (FWF) under grant SESAME (P17189-N04).

References

[1] Florin Cutzu, Riad Hammoud, and Alex Leykin. Estimating the photorealism of images: Distinguishing paintings from photographs. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II – 305–12, June 2003.

[2] R.P.W. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, and D.M.J. Tax. Prtools4 a matlab toolbox for pattern recognition, 2004.

[3] Qasim Iqbal and J.K. Aggarwal. Applying perceptual grouping to content-based image retrieval: building images. *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on*, 1, June 1999.

[4] Stefan Kuthan. Extraction of attributes, nature and context of images. Technical Report PRIP-TR-101, Vienna University of Technology, 2005.

[5] Jia Li and J.Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1075 – 1088, Sept. 2003.

[6] B. S. Manjunath and W.Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI - Special issue on Digital Libraries)*, 18(8):837–42, Aug 1996.

[7] Martin Szummer and Rosalind W. Picard. Indoor-outdoor image classification. In *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, pages 42 – 51, Jan. 1998.

[8] Aditya Vailaya, Mário A. T. Figueiredo, Anil K. Jain, and Hong-Jiang Zhang. Image classification for content-based indexing. *Image Processing, IEEE Transactions on*, 10(1):117 – 130, Jan. 2001.

[9] Aditya Vailaya, Anil K. Jain, and Hong-Jiang Zhang. On image classification: City vs. landscape. In *Content-Based Access of Image and Video Libraries, 1998. Proceedings. IEEE Workshop on*, pages 3 – 8, June 1998.

[10] Thomas Wagner. Texture analysis. *Handbook of Computer Vision and Applications, Signal Processing and Pattern Recognition*, 2:275–308, 1999.

[11] J.Z. Wang, Jia Li, and G. Wiederhold. Simplicity: semantics-sensitive integrated matching for picture libraries. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(9):947 – 963, Sept. 2001.