Conservation of Effort in Feature Selection for Image Annotation

Suzanne Little [1], Stefan Rüger [2]

[#] *Knowledge Media Institute, The Open University*
Walton Hall, Milton Keynes, United Kingdom
[1] s.little@open.ac.uk
[2] s.rueger@open.ac.uk

***Abstract*—This paper describes an evaluation of a number of subsets of features for the purpose of image annotation using a non-parametric density estimation algorithm (described in [1]). By applying some general recommendations from the literature and through evaluating a range of low-level visual feature configurations and subsets, we achieve an improvement in performance, measured by the mean average precision, from 0.2861 to 0.3800. We demonstrate the significant impact that the choice of visual or low-level features can have on an automatic image annotation system. There is often a large set of possible features that may be used and a corresponding large number of variables that can be configured or tuned for each feature in addition to other options for the annotation approach. Judicious and effective selection of features for image annotation is required to achieve the best performance with the least user design effort. We discuss the performance of the chosen feature subsets in comparison with previous results and propose some general recommendations observed from the work so far.**

I. Introduction

Much work has been expended on the topic of feature selection for machine learning, data analysis and information retrieval. When dealing with a very large and multi-dimensional data set, such as those found in multimedia or scientific applications, it is often critical to choose wisely the features used to index the set. The purposes of feature selection include reducing dimensionality, removing irrelevant and redundant information, reducing the amount of data needed and improving the accuracy of the annotator [2]. Feature selection assumes that from a set of available features there is an optimal subset that will be the most efficient and provide the best performance.

Automatic image annotation aims to reduce human effort in labelling and categorising images by training or otherwise configuring a system to classify images based on extracted low-level visual features such as colour, shape or texture. In this way image annotation systems attempt to find words such as "water", "building", "people" from an analysis of the image's pixels. The image annotation process generally consists of three phases: pre-processing, training (or classification or tuning) and application (or evaluation). In the *data pre-processing* phase raw image data is analysed, features are extracted and information is gathered. The output from this phase together with labels for the training set are passed to the *training* phase which applies some technique to develop a model to predict labels for previously unseen data. Some applications may apply a cross-validation step at this point to evaluate the performance, give feedback to the data pre-processing phase and repeat the training with altered configurations. The resulting model and a test set of unseen data are finally input to the *evaluation* phase which assesses the performance of the annotator.

In addition to the traditional uses of low-level features to annotate images with semantic labels, there is also potential benefit to be gained from applying image annotation to enhace query by example applications. Rather than comparing a query image using complex low-level feature descriptors it is possible to use the "keyword space" to index complex media objects using textual semantic labels and hence improve the performance and user experience [3]. In this way, even less accurate semantic labels are still incredibly useful as a single descriptor classifying visually similar images.

The annotation approach used in this evaluation is based on a nonparametric density estimation technique proposed by Yavlinsky et al. [1]. The nonparametric density estimator is applied to Bayes Theorem to provide an estimator of the true problem space density that makes no prior assumptions. This approach provides a simple, robust framework using global feature values for automated image annotation.

Features commonly used for automatic image annotation are generally visual features focused around colour, texture or shape descriptors which can be automatically extracted from an image. The values from these features may be continuous or discrete, numerical, a histogram or string in format and generally describe an image based on the identification of patterns or relationships at the pixel level. This level of description is less than ideal for a human user who would prefer higher-level semantic descriptions such as "tree", "bird" or "building". Image annotation attempts to bridge the well-known problem of the semantic gap [4].

Two main issues tend to complicate feature selection. The

MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.
978-1-4244-4464-9/09/$25.00

first is *redundancy* when combining top-performing features, which are strongly-related, has little or no added benefit. This is due to underlying similarity in the patterns which are described by the features. The second is *multivariate prediction* where high performing single features may not improve significantly when used in combination with other features while poorly performing features may demonstrate a exceptional improvement when applied together. These issues and their impact on performance are well illustrated by the evaluation described in section II-C.

The selection of features for image retrieval or indexing has been well analysed by Deselaers et al. [5] who describe a large variety of visual features and compare their performance quantitatively. The correlation of these features is also analysed and some of this information is applied to select features in our work. Some of the results obtained by Deselaers et al. are confirmed through our evaluation. The comprehensive overview of image retrieval written by Datta et al. [6] also covers the issues of feature selection and visual descriptors (or "image signatures") for identifying similar images. It is likely that the conclusions reached in these papers on image retrieval (where the goal is to find similar images or images that match a query) will apply equally well the image annotation tasks (where the goal is to find the most similar image and transfer the classification or label to the unknown image).

The work described in this article focuses on identifing suitable features at the data pre-processing stage before an annotator has been trained or applied. In contrast other approaches have used machine learning techniques to identify features or tune a classifier based on performance often using a wrapper approach [7] during the classification/training phase. Setia and Burkhardt [8] focus more on the configuration, tuning and weighting of a feature subset based on a quantitative computed measure of likelihood that describes the similarity of a feature and its discriminative ability. This is implemented using a wrapper approach with a support vector machine. A powerful but consuming approach is proposed by Lu et al. [9] who use a genetic algorithm to find the best features.

Incorporating feature subset selection feedback into the classification training phase of the annotation process helps to reduce the level of user configuration and therefore expertise required to set up an annotator for a particular dataset. This approach has been shown to work reasonably well in certain situations but may have less success when the dataset is high-dimensional and of small sample size due to the expense of evaluating every possible combination of features, variables and weights and the potential corruption of a good feature set by the presence of a "bad" feature. Kohavi et al. [7] and Viitaniemi and Laaksonen [10] also emphasise the influence and hence importance of choosing the evaluation metric to judge the performance improvements which is critical when performing feature subset selection and tuning.

The issue of feature selection is not limited to automatic image annotation. In a collection of short articles by multiple authors, Liu et al. [2] summarise many of the broader issues relating to feature selection in data mining and machine learning for a variety of application domains. Little et al. [11] used feature selection within a case-based classifier for biomedical data to weight (and hence select) features according to classification performance evaluated via cross-validation.

The motivation behind this work is two-fold: firstly to improve the performance and assess the stability of the non-parametric density estimation based (Npde) annotator. Secondly, to examine the level of effort (i.e., feature selection and fine-tuning) required to achieve significant results before diminishing returns reduce the value of the work.

The remainder of this paper discusses briefly the rationale behind the selection of features, the variables that can effect the choice and performance and presents the results from an evaluation. Some recommendations for feature selection in this area are proposed and conclusions presented.

II. Feature Selection and Evaluation

There are a wide variety of factors which influence the choice of features for automatic image annotation. Many of these are purely pragmatic factors such as the type of data, available analysis tools or space, size and efficiency concerns. Other decisions may require more indepth understanding of the data topography and the intended annotation approach. Even if an approach is used which conducts feature selection or weighting within the classification/training phase it is still likely that a subset of all available features will need to be chosen.

This section describes the features used and the evaluation of various combinations of features using a non-parametric density estimation approach [1] and the standard Corel5K image subset [12]. This subset consists of 5000 images from the Corel Stock Photo Library divided into a training set of 4500 with the remaining images used for testing. Images are labelled with 1–5 keywords from a vocabulary of 371 words. Only keywords with at least 2 images in the test set were evaluated which reduced the vocabulary to 179 keywords. While the Corel subset has been criticised for not providing sufficient variation for adequately assessing image annotation [13], it is still widely used and provides a very variable tool for comparison. We use the same setup as evaluations by [14], [15], [16], [17], [1] which differs slightly from that of the dataset's original paper.

In addition, previous experience has shown that results achieved using this dataset translate relatively to other larger and more complex datasets. Preliminary experiments conducted on a subset of the Getty image collection proposed by Yavlinsky et al. [1] confirm this.

A. Features

We used an internal tool (f_extract) to calculate feature files for each image and each feature. This package provides a large number of options to extract colour and texture features (among others) from images and calculate different descriptors based on statistical analysis of the histogram or by dividing the image into segments, weighting and combining the resulting values.

A conservative estimate, based on a subset of available features and configuration options in f_extract alone, finds in excess of 500 likely individual features that can be extracted and used for image annotation. The first problem is how to systematically define a likely subset of features. This section describes the colour and texture features used in this evaluation and gives the various configuration options available for each. These features represent commonly used and accepted features for image classification.

Colour: Colour is a key feature in identifying visual similarity and a number of colour space descriptors have been proposed often using different models of colour space based on definitions of human colour perception or printing needs. Functions for generating 3D colour histograms for an image are provided by f_extract. The available colour spaces used were: *RGB, HSV, HSL, Y'CbCr, CIELUV, CIELAB*. The number of bins that each axis of a colour space is divided into can also be configured. We used bins of 2+2+2; 4+2+2 and 8+8+8 in our initial evaluations.

Texture: Tamura (coarseness, contrast and directionality + window size and maximum range of coarseness), *Gabor* (scale and orientation)

Statistical Moments-based Features: Features themselves may be in a complex form, for example, as a vector of 100 numbers. The f_extract tool therefore provides another kind of feature which is the concatenation of three sets of statistics for the colour channel or texture histogram based on the first N statistical moments. In statistics, the moments provide an estimation of population parameters such as mean, variance, skewness, kurtosis, etc. The resulting feature for an image is the concatenation of these moment values and provides a simpler and potentially more meaningful summary of the feature than the complete vector.

Spatial awareness: The f_extract tool also enables spatial information from the image to be maintained by dividing the image up into regions. The required feature is extracted from each region independently and then the results are combined to produce a result which will be influenced by the distribution of the image. This can be done through specifying a tile size (e.g., T3x3 will divide the image into 9 equal segments) or through specifying a weighted spatial distribution type, either global (entire image) or focus, structured, local, centre for other arrangements. In addition to the feature specific options, each feature was also calculated over 3x3, 5x5 and 8x8 tiles plus global, local, focus and centred divisions of the image to incorporate some spatial information into the feature descriptor.

Once a likely subset of features has been determined, the second problem is to decide which features are likely to be redundant or highly correlated and which feature sets may suffer from multivariate prediction. Using information from Desalears et al. [5] and based on preliminary assessments, we developed a list of likely feature combinations. The general pattern used was to include one or more colour features plus optionally a Tamura texture feature and/or a Gabor texture feature. This list of feature subsets is given in Table VI.

Table I
SUMMARY OF RESULTS FOR INDIVIDUAL COLOUR FEATURES

Feature (# eval)	Mean	Median	Std Dev	Min/Max
all (176)	0.2074	0.2253	0.0614	0.0828 / 0.3120
stat. moments (168)	0.2127	0.2294	0.0576	0.0829 / 0.3120
spatial: any (144)	0.2236	0.2385	0.0545	0.0829 / 0.3120
spatial: 3x3 (24)	0.2569	0.2650	0.0260	0.1828 / 0.3016
spatial: 5x5 (24)	0.2648	0.2690	0.0324	0.1822 / 0.3120
spatial: 8x8 (24)	0.2376	0.2355	0.0231	0.1932 / 0.2747
spatial: Center (24)	0.2461	0.2450	0.0268	0.1651 / 0.2747
spatial: Focus (24)	0.1204	0.1189	0.0164	0.0829 / 0.1534
spatial: Local (24)	0.2202	0.2226	0.0252	0.1550 / 0.2576
bin dist: 2+2+2 (55)	0.2129	0.2306	0.0644	0.0829 / 0.3110
bin dist: 4+2+2 (55)	0.2149	0.2351	0.0609	0.0879 / 0.3120
bin dist: 8+8+8 (55)	0.2103	0.2214	0.0471	0.0999 / 0.2747
CIELAB (21)	0.2329	0.2520	0.0619	0.1245 / 0.3120
CIELUV (21)	0.2227	0.2411	0.0580	0.1176 / 0.2976
HSV (42)	0.2198	0.2391	0.0534	0.1119 / 0.2972
RGB (21)	0.1660	0.1817	0.0484	0.0829 / 0.2373

For the results given here, the distance metric is set to L_1-distance and features are unweighted. The summary results given in Table VII for Npde2 and Npde3 have had some feature weighting applied. More work is needed to fully explore the effect and meaning of distance measures and feature weighting in this context.

B. Procedure

The general process for evaluating a feature or feature subset was:

1) Analyse all the images using f_extract to extract the required feature descriptors.
2) Set the configuration options (feature list, distance metric, weights) for NpdeAnnotator.
3) Run NpdeAnnotator in evaluation mode which uses the training set to build up the aggregated model, queries for each label and calculates the mean average precision for each query (where more than two examples exist).
4) Save the results and basic timing information to file.
5) Remove temporary files for the trained model.
6) Repeat for the next feature subset.

This process was implemented in a shell script and executed over the list of individual features, the list of selected feature subsets and for the best performing feature combinations with a selection of varying feature weights.

C. Results and Discussion

Tables I, II and III summarise the performance of the individual colour, Gabor and Tamura features respectively, grouping the features according to the number of tiles, spatial weighting, histogram statistics, bin distribution and other feature specific configuration choices. Tables IV and V list the mean average precision of the top ten individual feature configurations.

The ordering of features in the tables is somewhat deceptive since there is, of course, no method for ordering features by increasing performance prior to evaluating them. However, it does demonstrate the relatively small changes that occur in

Table II
SUMMARY OF RESULTS FOR INDIVIDUAL GABOR FEATURES

Feature (#eval)	Mean	Median	Std Dev	Min/Max
All (36)	0.1725	0.1773	0.0313	0.0843 / 0.2067
Scale: 2 (12)	0.1562	0.1727	0.0407	0.0843 / 0.1948
Scale: 4 (12)	0.1825	0.1857	0.0246	0.1195 / 0.2067
Scale: 6 (12)	0.1788	0.1849	0.0200	0.1234 / 0.1975
Orientation: 2	0.1683	0.1815	0.0380	0.0843 / 0.2017
Orientation: 4	0.1754	0.1764	0.0294	0.0918 / 0.2067
Orientation: 6	0.1739	0.1745	0.0277	0.0933 / 0.2054
no spatial (9)	0.1346	0.1234	0.0405	0.0843 / 0.1862
spatial: 3x3 (9)	0.1925	0.1904	0.0108	0.1757 / 0.2067
spatial: 5x5 (9)	0.1900	0.1922	0.0097	0.1733 / 0.2014
spatial: 8x8 (9)	0.1729	0.1720	0.0041	0.1671 / 0.1792

Table III
SUMMARY OF INDIVIDUAL TAMURA FEATURES

Feature (# eval)	Mean	Median	Std Dev	Min/Max
all (320)	0.1193	0.1227	0.0263	0.0624 / 0.1761
¬ stat. moments (4)	0.0778	0.0753	0.0041	0.0743 / 0.0823
spatial: global	0.0974	0.0981	0.0071	0.0796 / 0.1153
spatial: 3x3 (44)	0.1419	0.1442	0.0169	0.0935 / 0.1749
spatial: 5x5 (44)	0.1429	0.1450	0.0213	0.1041 / 0.1761
spatial: 8x8 (44)	0.1230	0.1241	0.0163	0.0922 / 0.1525
spatial: center (44)	0.1341	0.1366	0.0111	0.1035 / 0.1521
spatial: focus (44)	0.0797	0.7880	0.0078	0.0624 / 0.1062
spatial: local (44)	0.1223	0.1230	0.0091	0.0979 / 0.1372
dist: 2+2+2 (104)	0.1265	0.1334	0.0288	0.0686 / 0.1761
dist: 4+2+2 (104)	0.1223	0.1288	0.0268	0.0675 / 0.1701
dist: 8+8+8 (104)	0.1112	0.1114	0.0192	0.0624 / 0.1521

Table IV
TOP TEN INDIVIDUAL FEATURES FOR *colour*: FEATURE, BIN DISTRIBUTION, TILE SIZE (ALL USING HISTOGRAM STATISTICAL MOMENTS)

Colour Feature	MAP
CIELAB, 4+2+2, 5x5	0.3120
CIELAB, 2+2+2, 5x5	0.3110
CIELAB, 4+2+2, 3x3	0.3016
CIELUV, 2+2+2, 5x5	0.2976
HSV (linear), 2+2+2, 5x5	0.2972
CIELUV, 4+2+2, 5x5	0.2956
Y'CbCr, 2+2+2, 5x5	0.2927
HSL (linear), 2+2+2, 5x5	0.2899
HSV (volume), 2+2+2, 5x5	0.2877
CIELUV, 4+2+2, 3x3	0.2862

Table V
TOP TEN INDIVIDUAL FEATURES FOR *Gabor*: SCALE, ORIENTATION, TILE SIZE AND *Tamura*: DISTRIBUTION, WINDOW SIZE, RANGE OF COARSENESS, TILE SIZE (USING STATISTICAL MOMENTS OF THE HISTOGRAM)

Gabor	MAP
4 4 3x3	0.2067
4 6 3x3	0.2054
4 2 3x3	0.2017
4 2 5x5	0.2014
6 2 5x5	0.1975
4 4 5x5	0.1974
6 4 3x3	0.1972
2 2 5x5	0.1948
4 6 5x5	0.1922
6 4 5x5	0.1915

Tamura	MAP
2+2+2 2 2 5x5	0.1761
2+2+2 2 2 3x3	0.1749
2+2+2 8 2 5x5	0.1732
2+2+2 6 2 5x5	0.1722
2+2+2 4 2 5x5	0.1719
4+2+2 2 2 5x5	0.1701
2+2+2 2 3 5x5	0.1677
2+2+2 2 3 3x3	0.1671
2+2+2 8 3 5x5	0.1661
4+2+2 4 2 5x5	0.1657

Table VI
TOP 15 BEST PERFORMING FEATURE SUBSETS. ALL COLOUR AND TAMURA FEATURES ARE DESCRIBED BY STATISTICAL MOMENTS OF THE HISTOGRAM, ALL BIN DISTRIBUTIONS ARE 2+2+2. THE VALUES FOR GABOR ARE FOR SCALE AND ORIENTATION. THE VALUES FOR TAMURA ARE THE WINDOW SIZE AND MAX RANGE OF COARSENESS. C INDICATES CENTER WEIGHTED SPATIAL DISTRIBUTION.

Feature subset	MAP
CIELAB.T3x3, HSV(volume).T3x3, gabor-4-4, Tamura-2-2-C	0.3648
CIELAB.T3x3, HSV(volume).T3x3, gabor-6-4, Tamura-2-2-C	0.3631
CIELAB.T3x3, HSV(linear)-C, gabor-6-4	0.3624
CIELAB.T3x3, HSV(linear)-C, gabor-4-4, Tamura-2-2-C	0.3617
CIELAB.T3x3, HSV(linear)-C, gabor-4-4	0.3600
CIELAB.T3x3, HSV(linear)-C, gabor-6-4, Tamura-2-2-C	0.3597
CIELAB.T3x3, HSV(linear)-C, gabor-4-4, Tamura-2-2.T3x3	0.3596
CIELAB.T3x3, HSV(volume).T3x3, gabor-4-4	0.3579
CIELAB.T3x3, HSV(linear)-C, gabor-6-4, Tamura-2-2.T3x3	0.3556
CIELAB.T3x3, HSV(volume).T3x3, gabor-6-4	0.3554
CIELAB.T3x3, HSV(volume).T3x3, gabor-6-4, Tamura-2-2.T3x3	0.3517
CIELAB.T3x3, HSV(volume).T3x3, gabor-4-4, Tamura-2-2.T3x3	0.3512
HSV(linear)-C, HSV(volume).T3x3, gabor-4-4, Tamura-2-2-C	0.3507
CIELAB.T3x3, HSV(linear)-C, Tamura-2-2.T3x3	0.3505
CIELAB.T3x3, HSV(linear)-C, gabor-4-4, Tamura-6-3.T3x3	0.3503

MAP for each feature and the very close performance of the highest performing features. The summary of the features, grouped by feature type and configuration options, gives a more general idea of the performance of an annotator when these options are varied. The general reliability of the best performing features is reassuring as it indicates the stability of the underlying approach and eliminates the chance that the distribution, window size, range of coarseness or tile size of the "best performing" configuration is merely an outlier for an otherwise poor annotator.

Table VI shows the fifteen best performing feature subsets and their MAP. The feature subsets were constructed from features in the top 20 performing individual features considering information from Desaelers et al. [5] about feature correlation, fitting features to a general pattern of combining 1 or more colour descriptors with 1 or more texture descriptors and, to a small extent, checking the time required to extract and process a feature compared with the potential improvement it offered.

The original feature set used by Yavlinsky et al. (Npde1) was a 3x3 tiled marginal histogram of global CIELAB colour space calculated across 2+2+2 bins and a 3x3 tiled marginal histogram of Tamura texture calculated across 2+2+2 bins with coherence of 6 and coarseness of 3. It applied euclidean distance (L_2-distance) for each feature value. The feature set and weights produced from the initial manual selection of features based on information from previous experiments (Npde2) used the same CIELAB and tamura features weighted as 1 and 0.25 respectively and added a Gabor texture descriptor with scale and orientation values of 4, weighted 0.5 and an extra colour feature described by a 3x3 tiled marginal global HSV histogram calculated on 2+2+2 bins and weighted 0.25. This configuration used L_1-distance. The final feature set and weights produced after a selected series of evaluations consisted of the same CIELAB, HSV and Tamura feature descriptors weighted 0.75, 0.5 and 0.5 respectively and a

Table VII
SUMMARY OF SYSTEM PERFORMANCE USING THE COREL5K DATASET (179 QUERY TERMS) AND SHOWING THE IMPROVEMENTS FOR THE NPDE USING THE GETTY DATASET. MAP=MEAN AVERAGE PRECISION; P%=PRECISION; R%=RECALL; N+=NUMBER OF WORDS WITH NON-ZERO RECALL; MBRM=MULTIPLE BERNOULLI REFERENCE MODEL [15] AND JEC=JOINT EQUAL CONTRIBUTION [18] (BOTH USING 260 QUERY TERMS).

Approach	**MAP**	**P%**	**R%**	**N+**	**MAP**(Getty)
Npde1 [1]	28.61	18	21	106	9.21
Npde2	37.15	21	19	93	11.86
Npde3	38.00	23	22	104	13.55
MBRM [15]	35	24	25	122	–
JEC [18]	–	27	32	139	–

Gabor texture feature using scale of 6 and orientation of 4 weighted 0.5. All distances were calculated using L_1-distance.

Table VII shows results from the original Npde annotator configuration plus the two new configurations (Npde2, Npde3) selected from the evaluation phase. In addition results from Feng et al. [15] using a Multiple Bernoulli Reference Model (MBRM) and from Makadia et al. [18] using a K-nearest neighbour, label transfer approach with Joint Equal Contribution (JEC) to combine the feature distances are shown. Also included are preliminary results from applying the same feature settings for the Npde annotator to a much more challenging subset of the Getty Image Archive website (described in [1]). Full analysis of results from this dataset is ongoing.

The improvement for Npde after applying some simple heuristics from [5] and increasing the number of features from 2 to 4 is strongly indicative of the influence of feature selection upon the performance of an automatic image annotator. The relatively small and insignificant gain achieved after more thorough competitive selection and tuning appears to indicate a plateau where further improvements do not result in significant performance gains.

The results for the JEC approach are extremely good. The authors note:

> "One reason for this exceptional performance may be due to the use of a wide spectrum of different features, contributing along different "orthogonal" factors" [18]

The seven features used for JEC were: colour – RGB, HSV & CIELAB and texture – Gabor (3 scales, 4 orientations), Haar Wavelet plus quantised versions of each. In addition the JEC approach used individual distance metrics selected for each feature rather than a common distance metric for all which may also contribute to the better performance. It is hoped to replicate this feature set and evaluate it in future work.

III. DISCUSSION AND RECOMMENDATIONS FOR FEATURE SELECTION

These general recommendations are based on the evaluations carried out so far on the Npde image annotation tool. The results achieved are consistent with those presented in other literature and demonstrate how feature selection can have a significant impact on the performance of an image annotation system. These recommendations are intended to assist in making judgments about selecting the most promising feature subsets for applications.

- If it is best to use only a single feature (e.g., for reasons of speed, space or computational complexity) then a colour feature such as CIELAB is likely to be the better choice.
- Using histogram statistics rather than the complete vector produces better results
- Using some spatial information (such as dividing the image into tiles and combining the output from each tile) produces better results.
- Combining a colour feature with a texture feature such as Gabor or Tamura improves the results.
- Increasing the number of features does not always give better results. More assessment is need to determine the most likely subset size but good results have been achieved with feature sets of 4 or less.

It seems reasonable that a colour feature based on human perception (such as CIELAB or CIELUV) will support better visual similarity results (however, not necessarily semantic similarity) and this appears to be consistent with both our results and that of other evaluations which support CIELAB as a valuable colour space descriptor. While most of the top performing feature sets contained a CIELAB feature (see table VI), it is interesting to note that an alternative feature set not using CIELAB but combining two very similar HSV features with Gabor and/or Tamura texture descriptors also achieve very high MAPs within 0.0140 to 0.0200 of the best performance. HSV based colour descriptors easily provide the next best individual performance after CIELAB.

The retention of some spatial information (or locally sensitive features) through the use of tiling or other weighted segmentation approaches generally improves the performance as shown in the individual feature tables I, II and III.

Applied individually, texture descriptors such as Gabor and Tamura do not perform as well as colour descriptors. However when applied in combination with a colour feature they result in a significant improvement in the overall MAP for the annotator. Choosing the best configuration options for the texture features is less clear as there is only slight differences in performance when values such as scale, orientation, coarseness and directionality are altered.

Overall, while it is tempting to focus on the slight improvements in the mean average precision, it is dangerous to place too much importance on improvements that are not significant enough to truly indicate a general better performance by the annotator. The detailed summary of the performance of individual features and subsets presented here is interesting to help identify those features, configurations and subsets which indicate promising performance by showing either significant improvements in the mean average precision for the dataset or confirming other indications about the stability of a feature (such as CIELAB) by consistently good average precisions and smaller deviations in performance.

IV. Conclusions and Future Work

We have demonstrated the importance of feature selection for image annotation and shown significant gains in MAP for the Npde annotator. The results and evaluation provided here demonstrate just how complex the selection of features can be and how many variables can potentially impact upon the performance of an image annotator. It is hoped that the general information here will be useful in determining the best path to take when choosing features for image annotation.

The adjustment of the feature set used by the Npde annotator has improved the MAP significantly from 0.2861 to a final result using weighted features of 0.3800. Preliminary evaluations using the more challenging Getty dataset also resulted in an improvement in the MAP from 9.21 to 13.55. It is also promising to see that many different features sets produce approximately comparable results within 0.0100 of the best performing combination. This indicates that the Npde annotator's performance is relatively stable and the top result is less likely to be an outlier that has been achieved through careful selection of features tuned specifically to the dataset.

This evaluation has demonstrated that visual features for image annotation are not independent. Two weakly performing features can provide significantly better performance when combined but equally a strongly performing feature can be negatively effected when combined with another feature. Some features are complimentary, some have no relationship, some are conflicting and some have strong correlation which renders their combination ineffective.

Choosing features based on general guidelines can provide results which are essentially equivalent in performance to features selected by expensive training and tuning. This can reduce development time and, hopefully, the issue of over fitting by selecting features based on a test set or through cross-validation. Given the expense of extracting some features from very large datasets and the consequent computation overhead required for calculating distances in the annotator, it is worthwhile considering the possible correlationships between features prior to training.

Finally, the evaluation and results produced so far give some promising avenues for further exploration. In the future we aim to apply this work to other data sets (specifically to a subset of the Getty photo collection [1] and the IAPR-TC12 collection from ImageCLEF [19]) to support our conjecture from preliminary experiments that our suggestions are generally valid across a wider selection of image types and further assess the impact on performance of feature pre-selection. In addition we intend to expand the feature set and investigate the influence of feature weighting and altering individually or globally applied distance metrics.

Acknowledgment

This work was funded in part by the European Union Sixth Framework Programme (FP6) through the integrated project PHAROS (IST-2006-045035).

References

[1] A. Yavlinsky, E. Schofield, and S. Rüger, "Automated image annotation using global features and robust nonparametric density estimation," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2005, pp. 507–517.

[2] H. Liu, E. Dougherty, J. Dy, K. Torkkola, E. Tuv, H. Peng, C. Ding, F. Long, M. Berens, L. Parsons, Z. Zhao, L. Yu, and G. Forman, "Evolving feature selection," *Intelligent Systems, IEEE*, vol. 20, no. 6, pp. 64–76, Nov.-Dec. 2005.

[3] J. Magalh aes, F. Ciravegna, and S. Rüger, "Exploring multimedia in a keyword space," in *MM '08: Proceeding of the 16th ACM international conference on Multimedia*. New York, NY, USA: ACM, 2008, pp. 101–110.

[4] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, 2000.

[5] T. Deselaers, D. Keysers, and H. Ney, "Features for image retrieval: an experimental comparison," *Information Retrieval*, vol. 11, no. 2, pp. 77–107, Apr. 2008.

[6] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, pp. 1–60, 2008.

[7] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1-2, pp. 273–324, 1997.

[8] L. Setia and H. Burkhardt, "Feature selection for automatic image annotation," in *Pattern Recognition*, 2006, pp. 294–303.

[9] J. Lu, T. Zhao, and Y. Zhang, "Feature selection based-on genetic algorithm for image annotation," *Knowledge-Based Systems*, vol. 21, no. 8, pp. 887 – 891, 2008.

[10] V. Viitaniemi and J. Laaksonen, "Evaluating performance of automatic image annotation: Example case by fusing global image features," in *Proceedings of the International Workshop on Content-Based Multimedia Indexing, 2007. CBMI '07.*, June 2007, pp. 251–258.

[11] S. Little, O. Salvetti, and P. Perner, "Evaluation of feature subset selection, feature weighting, and prototype selection for biomedical applications," in *Proceedings of the 9th European Conference on Case-Based Reasoning*, Trier, Germany, September 2008.

[12] P. Duygulu, K. Barnard, J. Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proceedings of the 7th European Conference on Computer Vision-Part IV*. Springer-Verlag, 2002, pp. 97–112.

[13] H. Müller, S. Marchand-Maillet, and T. Pun, "The truth about corel - evaluation in image retrieval," in *Image and Video Retrieval*, 2002, pp. 38–49.

[14] G. Carneiro and N. Vasconcelos, "Formulating semantic image annotation as a supervised learning problem," vol. 2. Los Alamitos, CA, USA: IEEE Computer Society, 2005, pp. 163–168.

[15] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2, pp. 1002–1009, June 2004.

[16] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. New York, NY, USA: ACM, 2003, pp. 119–126.

[17] J. Magalh aes and S. Rüger, "Information-theoretic semantic multimedia indexing," in *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*. New York, NY, USA: ACM, 2007, pp. 619–626.

[18] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *European Conference on Computer Vision, ECCV 2008*. Springer, 2008, pp. 316–329.

[19] M. Grubinger, "Analysis and evaluation of visual information systems performance," Ph.D. dissertation, School of Computer Science and Mathematics, Faculty of Health, Engineering and Science, Victoria University, Melbourne, Australia, 2007.