Image Metadata Estimation using Independent Component Analysis & Regression

M. Blighe, H. Le Borgne, N. O'Connor

Centre for Digital Video Processing,
Dublin City University, Ireland.

Abstract In this paper, we describe an approach to camera metadata estimation using regression based on Independent Component Analysis (ICA). Semantic scene classification of images using camera metadata related to capture conditions has had some success in the past. However, different makes and models of camera capture different types of metadata and this severely hampers the application of this kind of approach in real systems that consist of photos captured by many different users. We propose to address this issue by using regression to predict the missing metadata from observed data, thereby providing more complete (and hence more useful) metadata for the entire image corpus. The proposed approach uses an ICA based approach to regression.

1 Introduction

Semantic scene classification in image collections is currently a very active research area. The vast majority of approaches use low-level image features derived exclusively from scene content. Boutell et al [1] conducted an extensive survey on the state of the art in this area. They examined the features available, such as low-level features and camera metadata, and also provided a brief review of the learning and inference engines used for classification, e.g. k-nearest neighbour, Bayesian Classifier, etc.. They provided a review of scene classification systems and divided these systems into two types. Exemplar based systems use pattern recognition techniques using low-level image features or semantic features. Model-based approaches leverage the expected configuration of a scene. Interestingly, the use of camera metadata is not mentioned in any of the systems reviewed. The use of camera metadata, in combination with low-level features, is discussed in [2]. Here, a Bayesian network is used to fuse content-based data and metadata, with some promising results in specific contexts (e.g. indoor/outdoor classification).

One difficulty with such approaches is the non standard way in which digital cameras produce metadata. Although the Exchangeable Image File Format (EXIF) standard (http://www.exif.org) specifies the types of metadata available in digital cameras, different cameras from different manufacturers, produce different sets of metadata. This makes the use of image metadata in any scene classification system featuring content from multiple users extremely difficult. For example, the data used in our experiments is from the MediAssist collection [3] which currently consists of 11,203 images from 16 different users. Multiple models of digital camera were used in gathering this data and the metadata captured is not uniform across the test corpus.

In this paper, we demonstrate an approach to estimate the missing metadata, based on observed values for other images. In Section 2, we provide a high level description of the metadata used in the experiment. Section 3 describes our regression algorithm which is based on ICA. Section 4 describes the experiment we performed, whilst Section 5 outlines the results obtained. Conclusions are drawn in Section 6.

2 Digital Camera and System Metadata

The MediAssist collection of images used in our experiments consists of Exif metadata, as well as metadata specific to the MediAssist system. This includes GPS location information, captured using separate GPS devices when the photos were taken, location information downloaded from online gazetteers, local weather information, again downloaded after the event, user specific information, and manually annotated image information [4].

Not all of this metadata is relevant to our experiments and, in fact, the problem with missing information is only related to Exif information taken directly from digital cameras. However, the non Exif metadata is used in order to aid estimation of the missing data.

2.1 Exif Metadata

Exif is a specification for the image file format used by digital cameras. It is an open standard developed by the Japan Electronics Industry Development Association (JEIDA).

The metadata tags defined in the Exif standard cover a broad spectrum. It includes *date and time* information which digital cameras record and save in the metadata; *camera settings*, which includes static information such as the camera model and make, and information that varies with each image such as orientation, aperture, shutter speed, focal length; *location information*, which could come from a GPS receiver connected to the camera. Very few cameras currently support this, so a separate GPS device was used in the MediAssist project.

For our purposes, the information related to the camera settings and, therefore, the image capture conditions is the most relevant as most cameras will record the date/time information and location based information can be acquired after the photo has been captured. Examples of some of the metadata used can be seen in Table's 2 & 3 below.

2.2 Metadata Estimation

We created five databases of differing sizes and covering different metadata fields. Our main focus when creating these databases was to create a subset of the image collection containing a full set of metadata information, whilst focusing in particular on the Exif data. A lot of information available in the system was discounted at this point, as it was not particularly relevant to our experiments (e.g. manually annotated data). The final databases contained metadata broadly drawn from Exif information, location based GPS and gazetteer information, and user information.

The sizes of the databases, and the number of metadata fields available in each, are outlined in the table below. A number of fields were common to all databases

Table 1 Database Sizes

Database	No. of Images	No. Of Metadata fields
A	5635	24
B	5661	24
C	7010	23
D	4905	24
E	4869	27

(e.g. Exif Fnumber, Exif ExposureProgCode). However, certain fields were only available in one (or more) databases, but not all (e.g. Exif Brightness was only available in databases A and E, whilst Exif FlashValueCode was only available in databases B and C). Our approach (outlined below) was then used to estimate the missing metadata, based on information learned from the remaining images in the database. The metadata fields being estimated are manually removed from the test corpus, so the estimated results can be easily compared to the original values in order to evaluate the approach.

3 Estimation using ICA Regression

A detailed description of our overall system is provided in this section.

3.1 Independent Component Analysis

ICA is a statistical method which aims to express a set of observed variables as a linear combination of independent variables. Let us denote by $x = x_1, ..., x_n$ the observed variables, and likewise by $s = s_1, ..., s_n$ the independent underlying sources. Hence, the ICA model can be expressed as:

$$x = As \tag{1}$$

The difficulty of ICA is to estimate both the sources s and the linear mixtures A at the same time, knowing only the observed variable x. In [5], Comon showed it was possible assuming only statistical independence between the sources and, at most, only one Gaussian source. However, two ambiguities remain regarding the estimates. The first is that their magnitude is known give or take a scale factor. Note that this still leaves the ambiguity of the sign (we could multiply an independent component by -1 without affecting the model). The second ambiguity is that we cannot determine the order of the independent components, and a permutation of them will not change the result.

In [6] the authors note that the model in Eq. (1) is well-defined if, and only if, the components s_i are non-Gaussian. This is a fundamental requirement and is due to the fact that the sum of independent random variables has a distribution that is closer to Gaussian than any of the independent variables (according to the Central Limit Theorem). Thus, they can be used as measures of non-Gaussianity for ICA estimation. In the same paper they derive a fixed-point iteration scheme for ICA estimation called FastICA. The convergence of this algorithm is cubic (or at least quadratic), whilst other ICA algorithms based on gradient descent methods have only a linear convergence.

3.2 ICA Regression

Regression, in general, is the problem of estimating one variable given the values of some other variable or variables. In [7], Hyvarinen & Bingham propose to use ICA to achieve regression. They state the problem as follows. The variables of x are arranged so that the k

first variables form the vector of the observed variables $x_o = (x_1, ..., x_k)^T$, and the remaining variables form the vector of the missing variables $x_m = (x_{k+1}, ..., x_q)^T$. Thus the ICA model can be rewritten as

$$\begin{pmatrix} X_o \\ X_m \end{pmatrix} = \begin{pmatrix} A_o \\ A_m \end{pmatrix} s. \quad (2)$$

Hence, the regression problem is expressed in terms of the ICA model. The task is to predict x_m for a given observation of x_o. To be able to predict x_m, we must use (an estimate of) the joint probability distribution of x. Given that regression $\widehat{x}_m$ can be conventionally defined as the conditional expectation $E\{x_m|x_o\}$, Hyvarinen & Bingham propose the approximation below:

$$E\{x_m|x_o\} \approx A_m g(A_o^T x_o) \quad (3)$$

where the nonlinearity $g_i(u)$ equals the negative score function p_i'/p_i of the probability density of s_i plus an arbitrary linear term. For example, the tanh function is the score function of a mildly super-Gaussian distribution. So, the missing metadata x_m can be approximated using Eq. (3), where A_m and A_o are the mixing matrices obtained by learning and x_o represents the known dimensions of the testing data.

3.3 Estimation of the Nonlinearity

In the experiments of [7], the distribution of the independent components s_i are known. For example, the experimental data used in their experiments is generated according to distributions that are either strongly super-Gaussian, Laplacian or weakly super-Gaussian. Therefore, the nonlinearities given by the score functions can be easily calculated in each of these situations and used in the approximation to estimate x_m.

In our experiments, however, the distribution of the s_i are unknown. Therefore, the nonlinearity given by the score function is also unknown. We propose to use an identity function as the nonlinearity $g_i(u)$. This means that the approximation in Eq. (3) becomes:

$$E\{x_m|x_o\} \approx A_m(A_o^T x_o) \quad (4)$$

We can interpret the vector $A_o^T x_o$ as an initial linear estimate of s:

$$\widehat{s} = (A_o^T x_o) \quad (5)$$

Thus, the estimation of the missing metadata x_m is basically a linear reconstruction of the form $x_m = A_m \widehat{s}$.

4 Experiment

The data was first divided into two equally sized sets - one for learning and one for testing. The independent components were generated using the FastICA algorithm and the mixtures x were divided into observed x_o and missing x_m. The dimensionality of x_m was initially set to 1 or 2 in order to aid the visualisation of the results. In the preprocessing phase, the value of the missing variable x_m was first predicted by linear regression, and the residual of this regression was used in place of x_m in the remainder of the algorithm. After this linear prediction, the variables in x_o and x_m were whitened. The ICA estimation on the training data set gave the estimated values for the source signals s and the mixing matrix

$$A = \begin{pmatrix} A_o \\ A_m \end{pmatrix} \quad (6)$$

The test data set was used to compute estimates for the missing variable x_m. The value of the missing variable x_m was predicted using our approximation in Eq. (4). The success of the approximation is measured by the correlation coefficient between the true value of the missing variable and the value given by our approximation.

5 Results

The table below contains the best results obtained using our approximation. It should be noted that the results displayed above were obtained when only that particular variable was missing from the test data set (i.e. the dimensionality of x_m is set to 1). When the dimensionality of x_m was set to 2 or more, the performance of the algorithm detiorated rapidly.

Table 2 Exif fields with highest correlation coefficients

Metadata Name	Corr. Coef	Database
ExifBrightness	0.9802	*E*
ExposureValue	0.9693	*E*
ExifBrightness	0.9476	*A*
ExifExposureTime	0.9423	*E*
ExifFnumber	0.9146	*E*
ExifISOSpeed	0.9120	*E*
ExifShutterSpeed	0.8634	*E*
ExifFnumber	0.8350	*A*
ExifISOSpeed	0.8343	*A*
ExifFocalLengthNum	0.8246	*D*
ExifSubjectDistRangeCode	0.7706	*A*
ExifSubjectDistRangeCode	0.6657	*B*
ExifBrightness	0.6657	*D*

The best results are obtained with databases A & E, with Exif Brightness having the highest correlation coefficient in both cases. Other metadata fields with high correlation coefficients in both of these databases include Exif Fnumber and Exif ISOSpeed. It is clear, however, comparing databases A & E, that the results in database E are superior.

It should also be noted that the $ExposureValue$ obtained above is not an Exif datatype. This is a field generated in the MediAssist system and is a combination of shutter speed, aperture setting and ISO speed. It is also, therefore, a particularly useful field to be able to estimate.

Results obtained from the other databases were not as useful. For example, the highest correlation coefficient in database D is Exif FocalLengthNum with a correlation coefficient of 0.8246. Similarly for database B, the highest correlation coefficient obtained was 0.6657 for Exif SubjectDistanceRangeCode. In database C, the highest correlation coefficient obtained for a potentially useful Exif value was Exif Fnumber with a value of 0.5983.

However, it is worth noting that in all of the databases used in these experiments, high values for the correlation coefficient were obtained for various other metadata information in the database. However, none of this information was deemed to be useful in terms of a realistic practical application of metadata estimation. Table 4 below shows a selection of the metadata fields estimated with the highest correlation coefficient in each database. The high correlation coefficients obtained for these examples prove the viability of this approach to regression. High correlation coefficient scores were also obtained for many other fields in the database (e.g. Sunlight, Exif HeightInPixels, etc.), again showing the viability of this approach. So, although this information is not particularly interesting to estimate in the context of this application, as one would always expect to have these fields available, the algorithm may be useful in a different context on different data.

Table 3 Metadata fields with highest correlation coefficients

Metadata Name	Corr. Coef	Database
$ExifDateTaken$	0.9983	A
$LocalTimeMonth$	0.9965	A
$ExifDateTaken$	0.9998	B
$LocalDateTimeTaken$	0.9986	B
$ExifDateTaken$	0.9997	C
$LocalTimeMonth$	0.9991	C
$ExifDateTaken$	0.9998	D
$LocalDateTimeTaken$	0.9986	D
$LocalTimeMonth$	0.9990	E
$LocalDateTimeTaken$	0.9977	E

The fields in table 2 then, represent those which may be useful to predict in a realistic setting. It is also worth noting that, although Database E was the smallest in terms of the number of images, it contained the largest amount of metadata fields to use for training & testing.

Discrepancies between the results obtained from other databases are harder to interpret. Exif Brightness was not included in Databases B & C, so this may explain some of the poor performance in those databases (due to the fact that it had a high correlation coefficient in A & E). However, Database D contains Exif Brightness, and still had relatively poor correlation coefficient scores. A more detailed analysis is required to interpret the results further, but clearly the algorithm is very sensitive to the fields available for learning. In particular, database E contained the most Exif information, so it would appear that this information is critical in order to acquire good results.

6 Conclusions and Future Work

This paper has demonstrated the potential of the ICA based approach to regression. The high correlation coefficients obtained for certain Exif data fields prove the viability of the approach, however work needs to be done to improve the system into one which could be used in a realistic setting. The fact that other fields can be estimated accurately, however, demonstrates that this general approach may be useful in different application settings.

Future work will focus on the refinement of the algorithm outlined in this work, in order to improve results. The estimation of the nonlinearity $g_i(u)$ used in Eq. (3) may be improved by using a non-parametric approach to density estimation. This should improve the obtained results. Another approach under consideration is the implementation of a numerical integration approach to estimating the missing variables.

Acknowledgement: This material is based on works supported by the Entreprise Ireland funded MediAssist project, as well as the aceMedia Project (acemedia.org), part of the EU 6th Framework Program.

References

1. M. Boutell and C. Brown. Review of the state of the art in semantic scene classification. Technical report, Department of Computer Science, University of Rochester, Rochester, NY, 2002.
2. M. Boutell and J. Luo. Beyond pixels: Exploiting camera metadata for photo classification. *IEEE Conference on Computer Vision and Pattern Recognition*, June 2004.
3. N. Murphy C. Gurrin and G. Jones. Mediassist: Managing personal digital photo archives. ERCIM News, July 2005. No. 62.
4. N. O'Hare G. Jones, C. Gurrin and A.F. Smeaton. Combination of content analysis and context features for digital photograph retrieval. *IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, London*, 2005.
5. P. Comon. Independent component analysis: A new concept? *Signal Processing*, 36:287–314, 1994.
6. A. Hyvarinen E. Oja and J. Karhunen. *Independent Component Analysis*. John Wiley & Sons, 2001.
7. A. Hyvarinen and E. Bingham. Connection between multilayer perceptrons and regression using independent component analysis. *Elsevier Science Neurocomputing*, 50:211–222, 2003.