Search Optimization for JPEG Quantization Tables

using a Decision Tree Learning Approach

Sharon Gieske
6167667

Supervisors: Zeno Geradts (NFI)

Master System and Network Engineering
University of Amsterdam

2014-07-02

Table of Contents

Motivation

- Growing popularity for taking pictures
- Digital images often recovered in forensic investigations
- Identify origin of images to a specific camera or common source
- Large sets of images are retrieved

Camera Identification:

- Intrinsic features of camera hardware give more reliable results[2]
- Sensor Imperfections, CFA Interpolation, Image Features

JPEG quantization tables

JPEG compression:

- RGB to Luminance-Chrominance colour space
- Splitting into two 8×8 blocks
- Discrete Cosine Transform (spatial domain → frequency domain)
- Compression ratio
- Correlated to camera make/model

'..is reasonably effective at narrowing the source of an image to a single camera make and model or to a small set of possible cameras.'[1]

Decision tree learning algorithm

Camera identification problem → pattern recognition problem:

- map feature set to corresponding label

Decision tree learning algorithm:

- Rule based, generates best splits
- Simple to interpret / human readable

Research Question

Can searching through JPEG quantization tables be optimized with the use of decision tree learning?

Subquestions:

1. Can identifiable parameters be found in JPEG quantization tables?
2. What is the performance of decision tree learning with JPEG quantization tables?

Overview

1. Extract quantization tables from images
2. Generate feature set
3. Train decision tree classifier (make/model)
4. Evaluate classifications
5. Compare against method using hash database

Data Preprocessing and Training

1. **Extract quantization tables from images**
   - Unix command: djpeg
2. **Generate feature set**
   - Add features: sum, min, max, mean, median, var, std
   - Run feature selection
3. **Train decision tree classifier**
   - CART: combines classification and regression trees

Evaluation

4. **Evaluate with weighted F$_\beta$-score**
    - Recall is more important: $\beta = 2$

$$F_\beta = 1 + \beta^2 * \frac{precision * recall}{(\beta^2 * precision) + recall} \quad (1)$$

5. **Compare against method using hash database**
    - Database of hashed quantization tables
        - 1→1 mapping
        - 1→n mapping
    - Use same training and validation data

Results

Dataset:

- 45,666 images (NFI & Dresden Image Database)
- 41 camera models
- 19 camera makes
- 1,016 unique quantization tables

Identifiable parameters: 50 out of 128
603 nodes, depth of 26

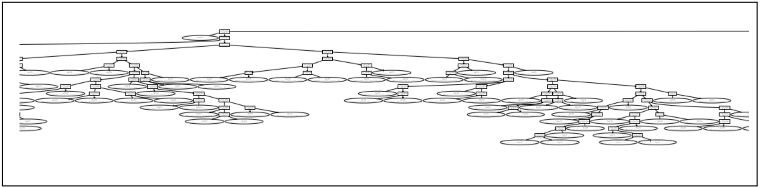

Figure: Partial Decision Tree

Zoom in: F2-score for camera make

Make	F2		Make	F2
Kodak	99 %		Praktica	43 %
Ricoh	94 %		Nikon	86 %
Panasonic	79 %		Casio	99 %
PS	100 %		Canon	98 %
Olympus	64 %		Logitech	100 %
Sony	58 %		Motorola	100 %
Agfa	78 %		Epson	100 %
Rollei	84 %		BlackBerry	100 %
Samsung	67 %		Pentax	80 %
FujiFilm	96 %			

Table: F2-score for camera make

Decision tree vs Hash databases

- 5-Fold Stratified Cross Validation
- 80 % Train set, 20 % Validation set

Algorithm	Precision	Recall	F2-score
Hash (1-1)	79 %	68 %	68 %
Hash (1-n)	50 %	99 %	83 %
Decision tree	90 %	89 %	89 %

Table: Camera Make Identification

Algorithm	Precision	Recall	F2-score
Hash (1-1)	54 %	39 %	37 %
Hash (1-n)	50 %	98 %	83 %
Decision tree	78 %	82 %	80 %

Table: Camera Model Identification

Discussion

- Both methods are prone for overfitting
- Hash database holds larger search space
- Training hash database is quicker

Conclusions

- Parameters can be reduced to 50
- Decision tree classifier gains better F2-score of 89% (make)
- 1→N hash database gains better F2-score of 83% (model)

- Decision tree classifier is more flexible, reduces search space, but harder to train than 1→N hash database

Future work:

- Compare to other learning algorithms
  - Naive Bayes
- Extend feature set

Questions?

References I

Hany Farid.
Digital image ballistics from jpeg quantization.
Technical report, Dartmouth College, Department of Computer Science, 2006.

Tran Van Lanh, Kai-Sen Chong, Sabu Emmanuel, and Mohan S Kankanhalli.
A survey on digital camera image forensic methods.
In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 16–19. IEEE, 2007.