Towards Open-Set Forensic Source Grouping on JPEG Header Information

Patrick Mullan, Christian Riess, Felix Freiling

DFRWS EU 2020

Multimedia Security Group
Department of Computer Science
Friedrich-Alexander University Erlangen-Nürnberg (FAU)

Goals of digital multimedia forensic

Research question:

"Given an image, what can we know about its origins?"

images from:
de.wikipedia.org/wiki/Canon_EOS_450D
de.wikipedia.org/wiki/IPhone_5s
de.wikipedia.org/wiki/Canon_PowerShot_S

Goals of digital multimedia forensic

Research question:

"Given an image, what can we know about its origins?"

One important goal of multimedia security:

Reconstruct image history, e.g., it's provenance

images from:
de.wikipedia.org/wiki/Canon_EOS_450D
de.wikipedia.org/wiki/IPhone_5s
de.wikipedia.org/wiki/Canon_PowerShot_S

Goals of digital multimedia forensic

Research question:

"Given an image, what can we know about its origins?"

One important goal of multimedia security:

Reconstruct image history, e.g., it's provenance

- Identify processing software of multimedia content

images from:
de.wikipedia.org/wiki/Canon_EOS_450D
de.wikipedia.org/wiki/IPhone_5s
de.wikipedia.org/wiki/Canon_PowerShot_S

Goals of digital multimedia forensic

Research question:

"Given an image, what can we know about its origins?"

One important goal of multimedia security:

Reconstruct image history, e.g., it's provenance

- Identify processing software of multimedia content
- Identify paths the content was distributed over

images from:
de.wikipedia.org/wiki/Canon_EOS_450D
de.wikipedia.org/wiki/IPhone_5s
de.wikipedia.org/wiki/Canon_PowerShot_S

Goals of digital multimedia forensic

Research question:

"Given an image, what can we know about its origins?"

One important goal of multimedia security:

Reconstruct image history, e.g., it's provenance

- Identify processing software of multimedia content
- Identify paths the content was distributed over
- Identify source of multimedia content

images from:
de.wikipedia.org/wiki/Canon_EOS_450D
de.wikipedia.org/wiki/IPhone_5s
de.wikipedia.org/wiki/Canon_PowerShot_S

Goals of digital multimedia forensic

Research question:

"Given an image, what can we know about its origins?"

One important goal of multimedia security:

Reconstruct image history, e.g., it's provenance

- Identify processing software of multimedia content
- Identify paths the content was distributed over
- Identify source of multimedia content

This work

images from:
de.wikipedia.org/wiki/Canon_EOS_450D
de.wikipedia.org/wiki/IPhone_5s
de.wikipedia.org/wiki/Canon_PowerShot_S

Use cases of investigating image provenance

- Police works:

  Sift through tons of images on a confiscated hard drive and find those images related to a case

Use cases of investigating image provenance

- Police works:

  Sift through tons of images on a confiscated hard drive and find those images related to a case

- Insurance companies:

  Do the images sent in from the client really show the claimed damage, or are the images edited/copied together?

Use cases of investigating image provenance

- Police works:

  Sift through tons of images on a confiscated hard drive and find those images related to a case

- Insurance companies:

  Do the images sent in from the client really show the claimed damage, or are the images edited/copied together?

- Social media sites:

  Are the uploaded images by the user, or subject to copyright-infringement?

Use cases of investigating image provenance

- Police works:

  Sift through tons of images on a confiscated hard drive and find those images related to a case

- Insurance companies:

  Do the images sent in from the client really show the claimed damage, or are the images edited/copied together?

- Social media sites:

  Are the uploaded images by the user, or subject to copyright-infringement?

- …

Use cases of investigating image provenance

- Police works:

  Sift through tons of images on a confiscated hard drive and find those images related to a case

- Insurance companies:

  Do the images sent in from the client really show the claimed damage, or are the images edited/copied together?

- Social media sites:

  Are the uploaded images by the user, or subject to copyright-infringement?

- … ⇒Examples require automatized, scalable, and fast solutions

This talk is about

Towards Open-Set Forensic Source Grouping on JPEG Header Information

This talk is about

Towards Open-Set Forensic Source Grouping on JPEG Header Information

Next slides give background information on

- (Digital) images, and where to find cues for forensic
- What counts as source; a hierarchical overview

This talk is about

Towards Open-Set Forensic Source Grouping on JPEG Header Information

Next slides give background information on

- (Digital) images, and where to find cues for forensic
- What counts as source; a hierarchical overview

- Define "open-set" problem
- Illustrate a method with experiments to solve the open-set problem

What is a (digital) image?

Image – visual content, produced in photography

What is a (digital) image?

Image – visual content, produced in photography

Technical:

What is a (digital) image?

Image – visual content, produced in photography

Technical:

- Represent content as:
  - Compressed bytestream in a data file
  - Uncompressed data in memory (e.g. display on screen)

What is a (digital) image?

Image – visual content, produced in photography

Technical:

- Represent content as:
  - Compressed bytestream in a data file
  - Uncompressed data in memory (e.g. display on screen)
- Content is put in an image file, here JPEG :
  - Optional, additional, information, e.g., metadata
  - Further header information defining semantic of bytestream
  - Contains bytestream

0xFF 0xD8 …
Image resolution, … Pixel density, … EXIF metadata, …
Scan format, … Huffman tables, … Quantization tables, …
Encoded image, … 0xFF 0xDA … … 0xFF 0xD9

What is a (digital) image?

Image – visual content, produced in photography

Technical:

- Represent content as:
  - Compressed bytestream in a data file
  - Uncompressed data in memory (e.g. display on screen)
- Content is put in an image file, here JPEG :
  - Optional, additional, information, e.g., metadata
  - Further header information defining semantic of bytestream
  - Contains bytestream

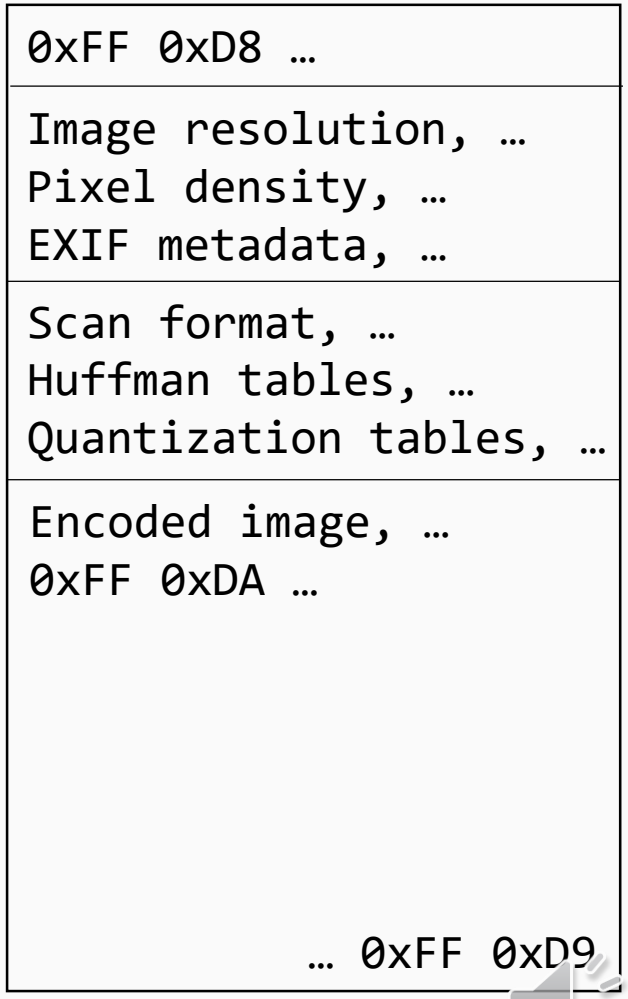
```
0xFF 0xD8 …

Image resolution, …
Pixel density, …
EXIF metadata, …

Scan format, …
Huffman tables, …
Quantization tables, …

Encoded image, …
0xFF 0xDA …

… 0xFF 0xD9
```

What is a (digital) image?

Image – visual content, produced in photography

Technical:

- Represent content as:
  - Compressed bytestream in a data file
  - Uncompressed data in memory (e.g. display on screen)
- Content is put in an image file, here JPEG :
  - Optional, additional, information, e.g., metadata
  - Further header information defining semantic of bytestream
  - Contains bytestream

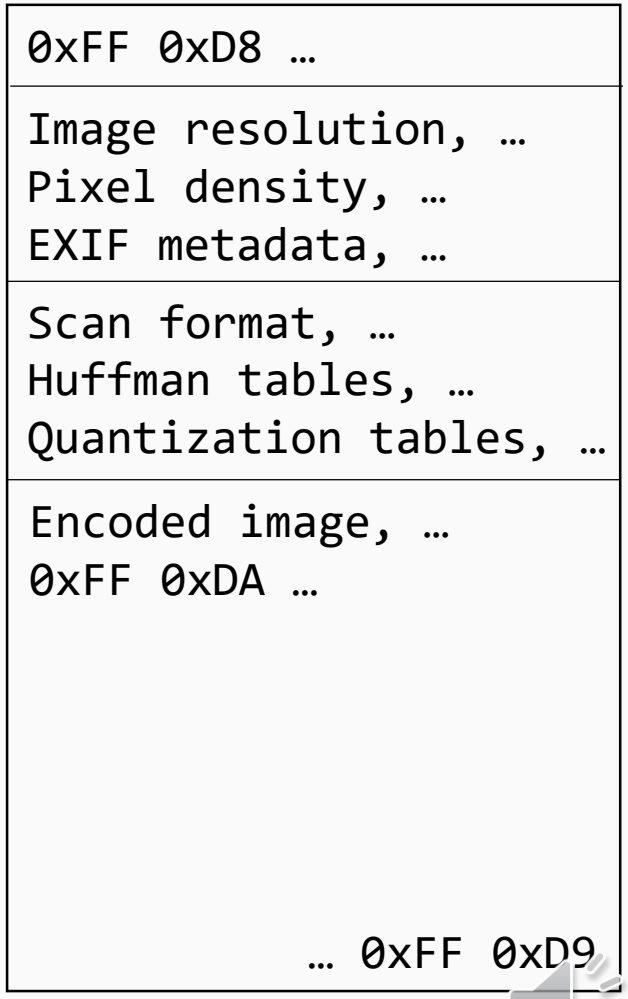


What is a (digital) image?

Image – visual content, produced in photography

Technical:

- Represent content as:
  - Compressed bytestream in a data file
  - Uncompressed data in memory (e.g. display on screen)
- Content is put in an image file, here JPEG :
  - Optional, additional, information, e.g., metadata
  - Further header information defining semantic of bytestream
  - Contains bytestream

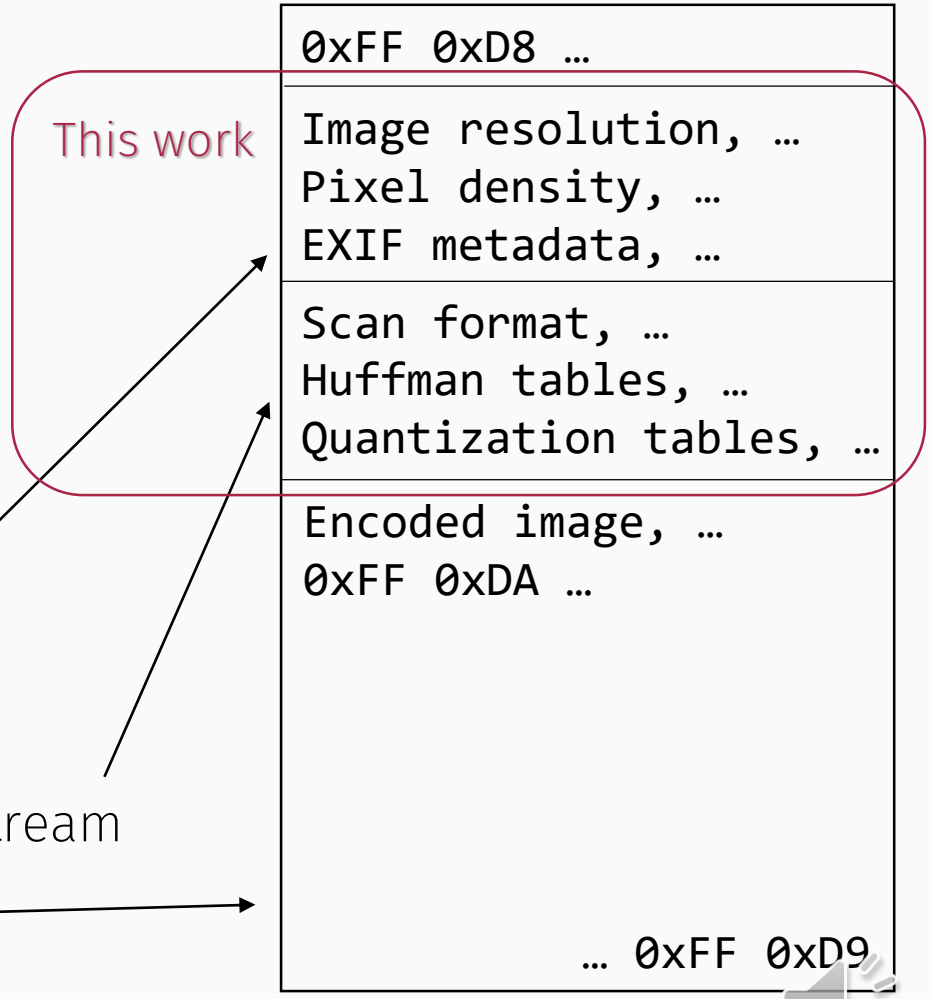


What is a (digital) image?

Image – visual content, produced in photography

Technical:

- Represent content as:
  - Compressed bytestream in a data file
  - Uncompressed data in memory (e.g. display on screen)
- Content is put in an image file, here JPEG :
  - Optional, additional, information, e.g., metadata
  - Further header information defining semantic of bytestream
  - Contains bytestream

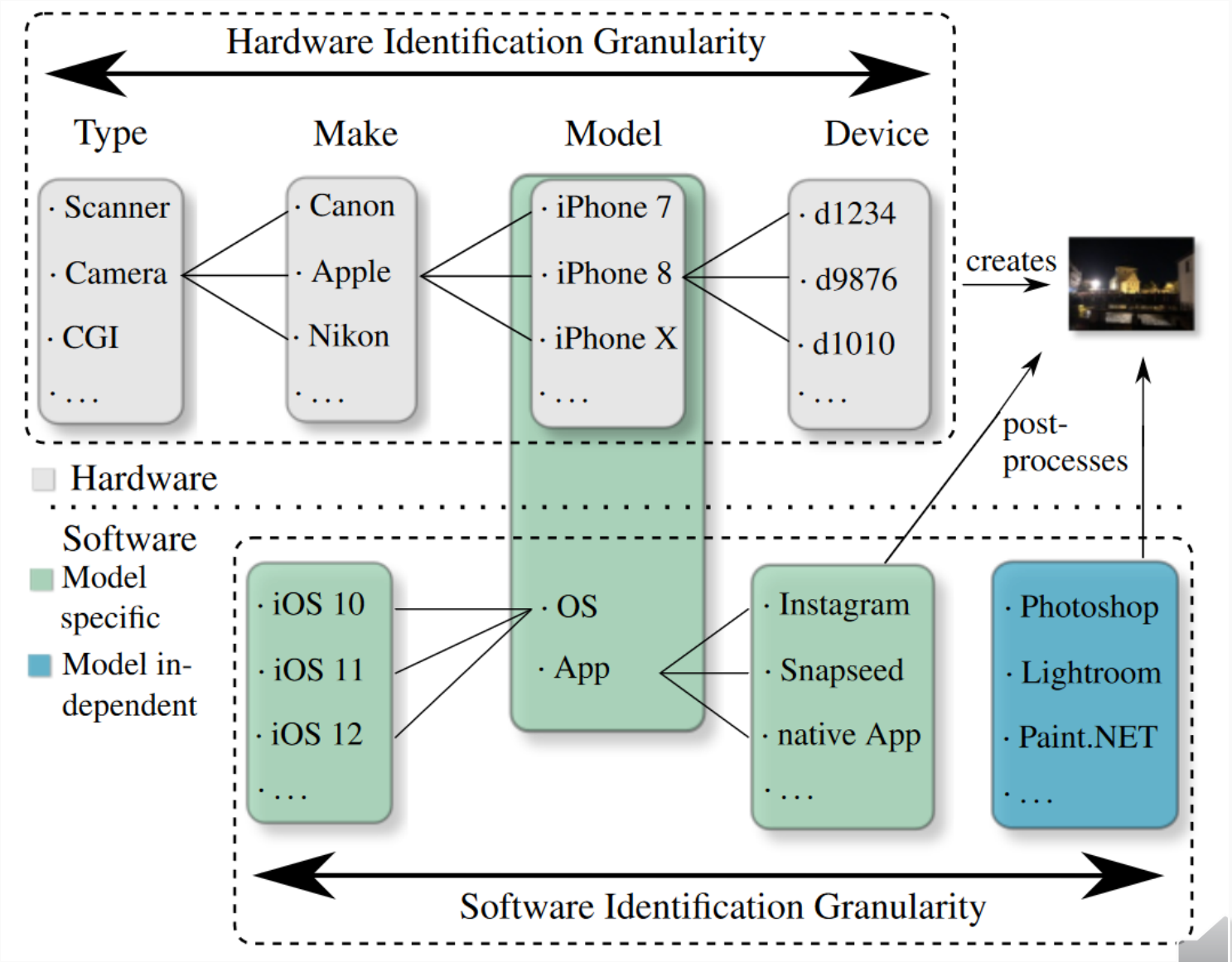


What is source identification?

Different source granularities:

Top: Granularity on hardware level

Bot.: Granularity on software level



Two starting points for doing source identification

Two starting points for doing source identification

- Pixel information – analyze the decoded image
  - A lot of literature focuses on that
  - Machine learning (e.g., deep learning nowadays) to predict the source of query image

Two starting points for doing source identification

- Pixel information – analyze the decoded image
  - A lot of literature focuses on that
  - Machine learning (e.g., deep learning nowadays) to predict the source of query image
- Header information – analyze the content and composition of the image file
  - The image file is always at hand:
    - Hence offers a natural entry point for an investigation
    - Offers information on a coarse grouping
  - Further advantages over pixel based methods:
    - Relatively easy to apply/implement
    - Lightweight operations, hence really fast methods

Two starting points for doing source identification

- Pixel information – analyze the decoded image
  - A lot of literature focuses on that
  - Machine learning (e.g., deep learning nowadays) to predict the source of query image
- Header information – analyze the content and composition of the image file
  - The image file is always at hand:
    - Hence offers a natural entry point for an investigation
    - Offers information on a coarse grouping
  - Further advantages over pixel based methods:
    - Relatively easy to apply/implement
    - Lightweight operations, hence really fast methods

This work

Proposal to the open-set problem

- General problem in source identification: Closed-set knowledge

  Ongoing development of camera models, e.g., new models are released nearly daily

  -> Keeping a database of all models up-to-date can be considered as intractable

Proposal to the open-set problem

General problem in source identification: Closed-set knowledge

Ongoing development of camera models, e.g., new models are released nearly daily

-> Keeping a database of all models up-to-date can be considered as intractable

Proposal to the open-set problem

General problem in source identification: Closed-set knowledge

Ongoing development of camera models, e.g., new models are released nearly daily

-> Keeping a database of all models up-to-date can be considered as intractable

We propose to formulate source identification as:

- a open-set problem, with
- solvable with classification frameworks,
- with the goal of predicting the make of a previously unseen model.

Data

- We collected 2,833,349 images from Flickr
- Parsing of metadata in header: EXIF:Make and EXIF:Model:
  - Normalization: (Samsung, SAMSUNG, Samsung techwin co., ltd, ...,)
  - Cleaning: malformed strings (cAnon), or contradicting entries (iPhone from Canon)
- Sort images into presumably unedited and edited images

Canon

EXIF:Software	Frequency
--	286,173
Digital Photo Professional	14,772
Picasa	11,250
Adobe Photoshop CS3 Windows	9,030

Nikon

EXIF:Software	Frequency
Ver.1.00	60,633
Ver.1.01	33,435
Ver.1.10	11,057
Ver.1.03	7,580

Studied feature set

- We use two types of feature sets in our work:
  - Quantization tables as feature set
  - And count of entries in each IFD -> Histogram over the IFDs
- EXIF-Standard suggests logic groups, called Image file directories (IFDs)

IFD group name	Meaning
IFD0	Metadata concerning main image
IDF1	Metadata concerning thumbnail
ExifIFD	Photometric information
GPS	Geolocation
MakerNotes	Proprietary information
ICC Profile	Color information (not part of EXIF Standard)

Roadmap to develop an open-set system

- Characterization of the variability of header information:

  Intra-make similarities and inter-make differences

Roadmap to develop an open-set system

- Characterization of the variability of header information:

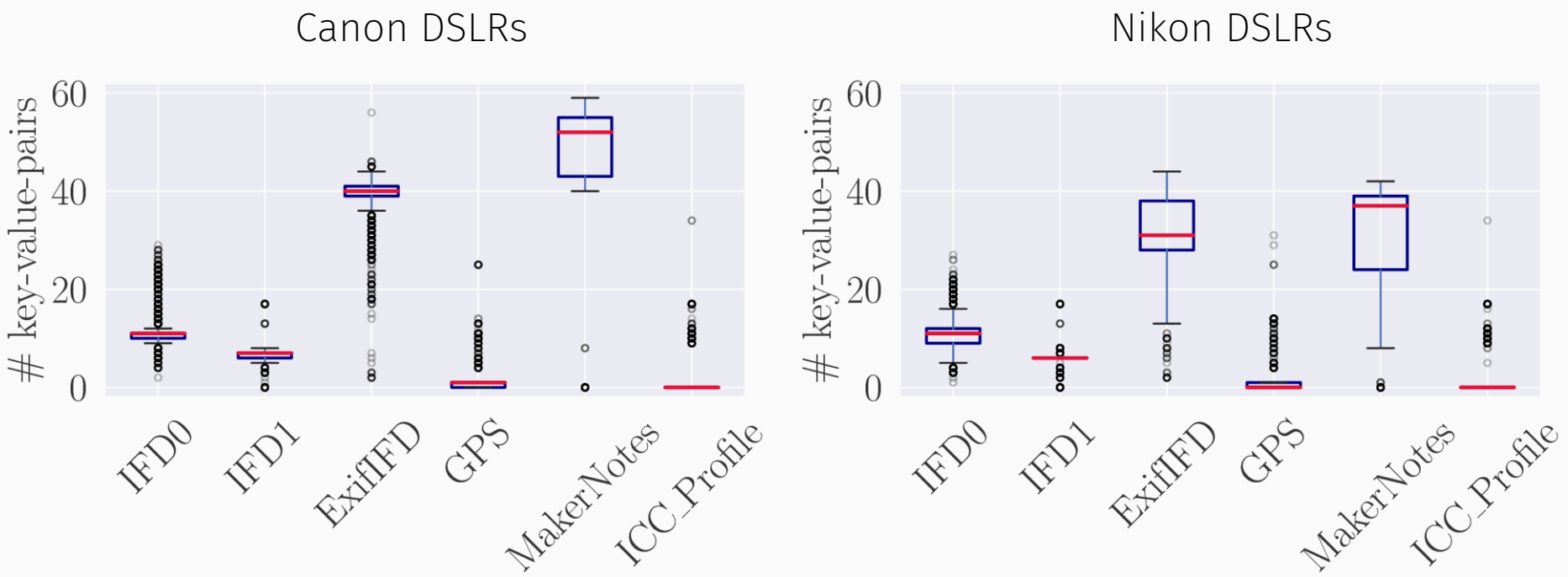
  Intra-make similarities and inter-make differences

- Experimental evaluation:

  - How well the make of a camera can be predicted on images from new, unknown models

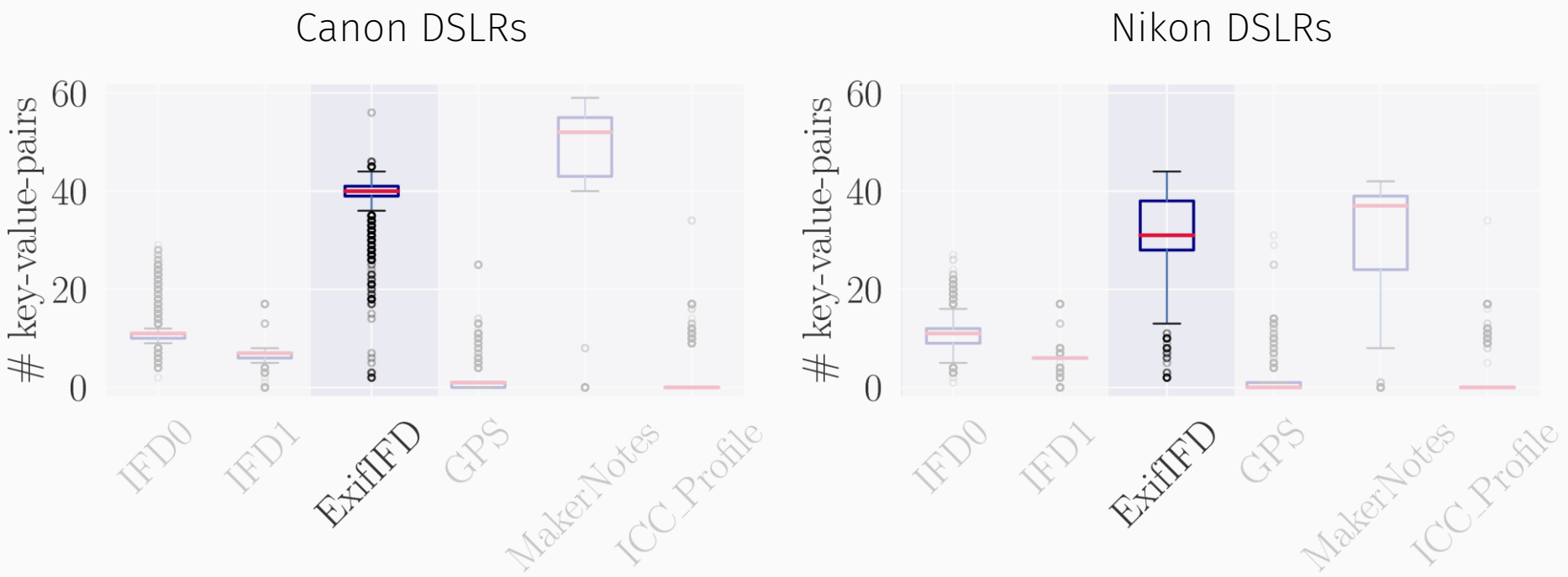
  - How well is this prediction on images that were processed by additional software

Characterization of the variability of header information



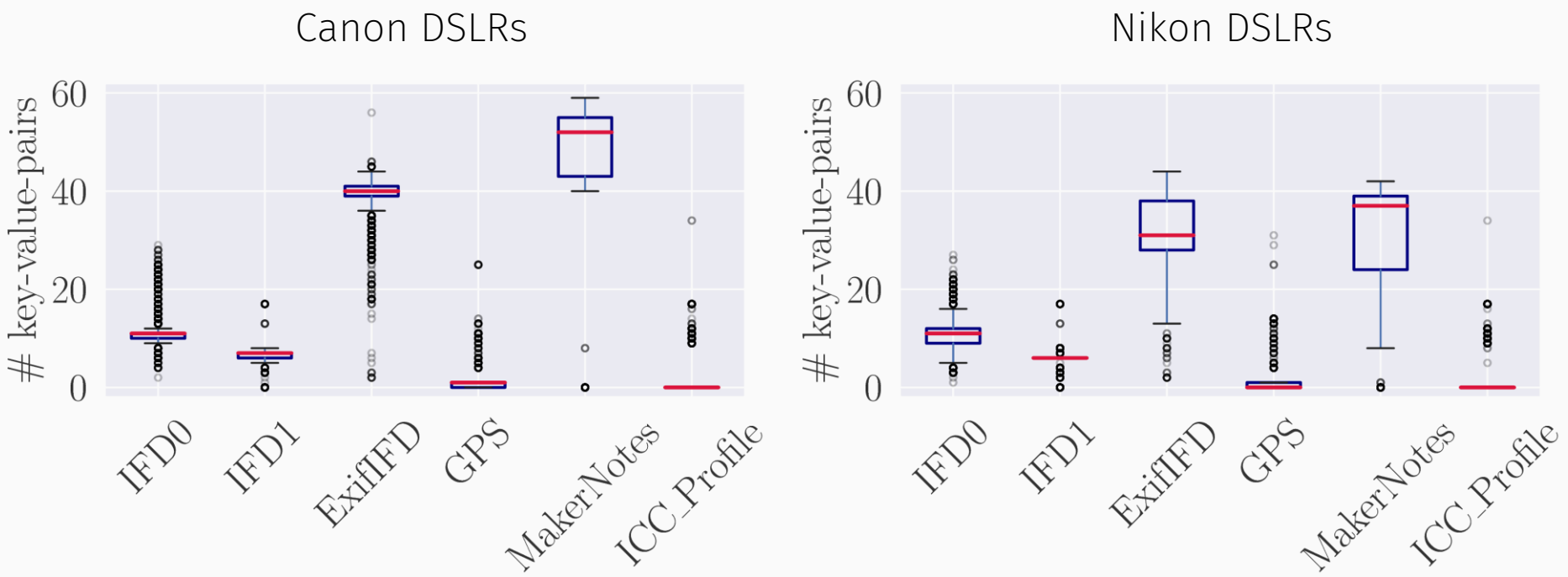
- Manufacturer have freedom to populate IFDs with any number of entries
- The histograms over IFDs seem to be characteristic for manufacturers

Characterization of the variability of header information



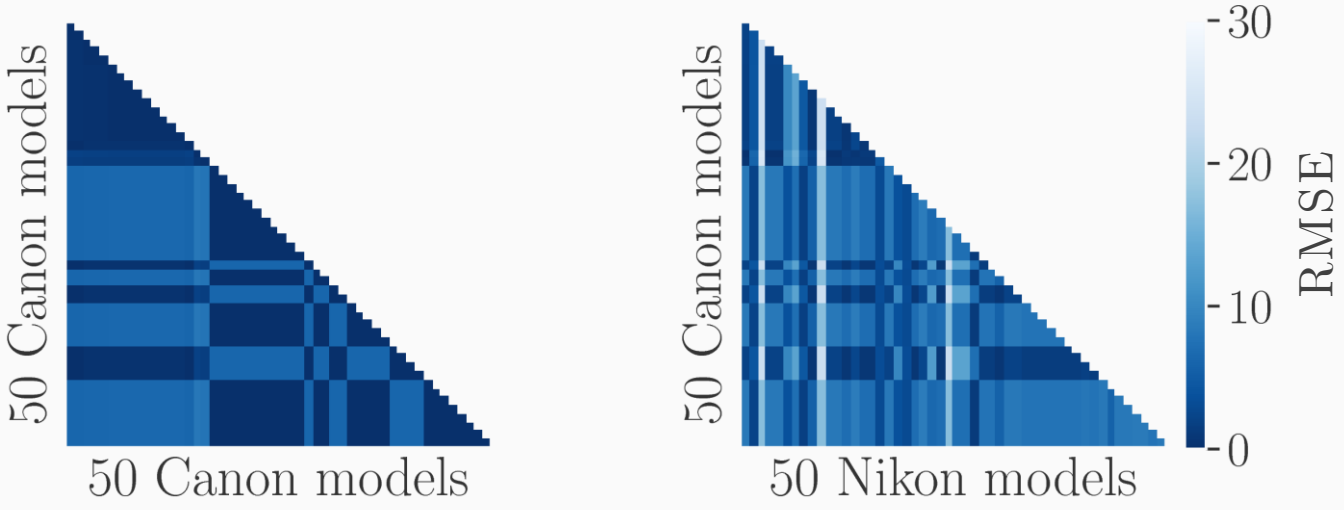
- Manufacturer have freedom to populate IFDs with any number of entries
- The histograms over IFDs seem to be characteristic for manufacturers

Characterization of the variability of header information



- Manufacturer have freedom to populate IFDs with any number of entries
- The histograms over IFDs seem to be characteristic for manufacturers

Similarity among quantization matrices between and across makes



- Each axis lists 50 models of cameras
- Adjacency matrices show root mean squared error (RMSE) among typical quantization tables of those models, where stronger saturation means bigger similarity

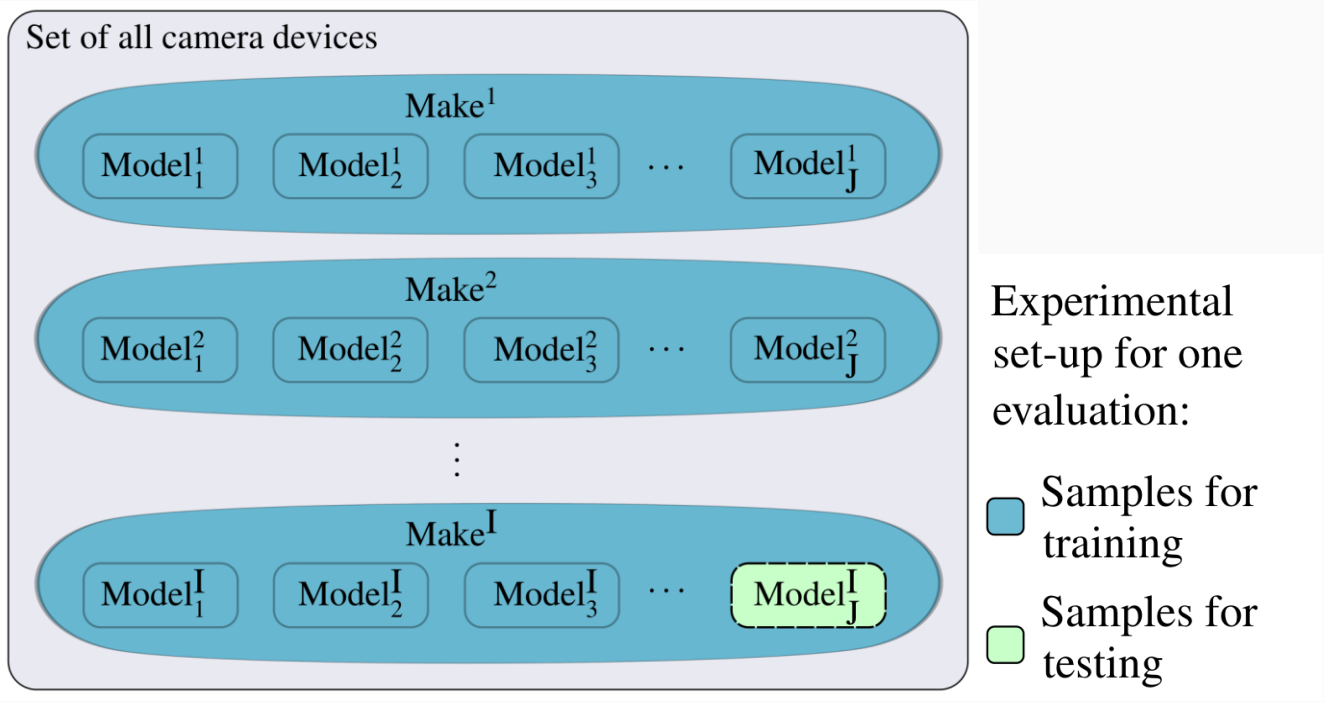
-> Manufacturers reuse quantization tables for different models

-> Many quantization tables seem to be unique to Nikon or Canon

Experiments

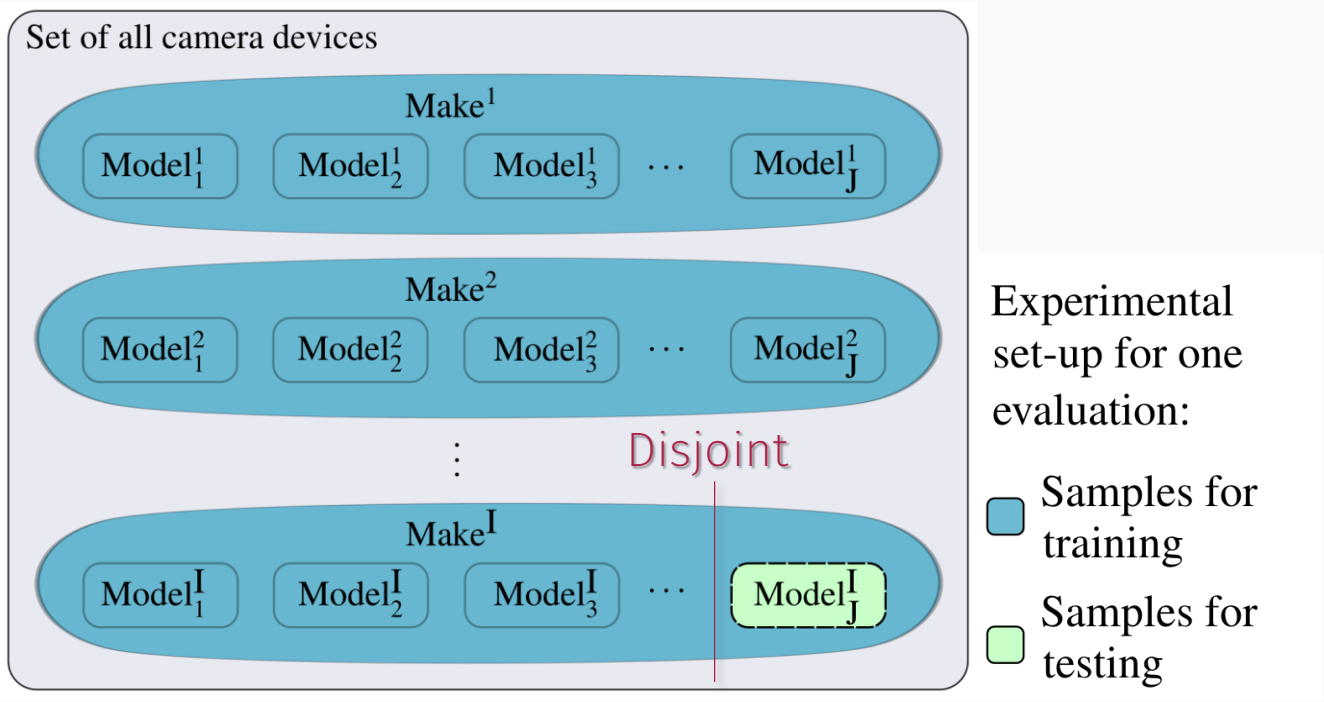
1. Predict make of images that stem from new, unknown, models
2. Predict make of images that were processed with additional software

Set-up for experiment on associating unknown model with its make



- Learn a classifier, that predicts the make of an input image
- The training data does **not** have representatives of a **specific model**
- The test data are images from this model, that was not available at training

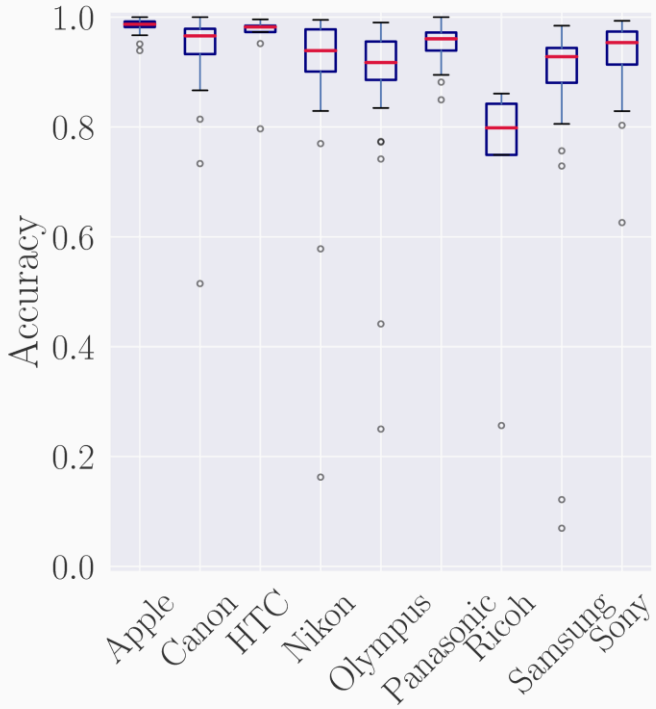
Set-up for experiment on associating unknown model with its make



- Learn a classifier, that predicts the make of an input image
- The training data does **not** have representatives of a **specific model**
- The test data are images from this model, that was not available at training

Classification performance, considering different feature sets

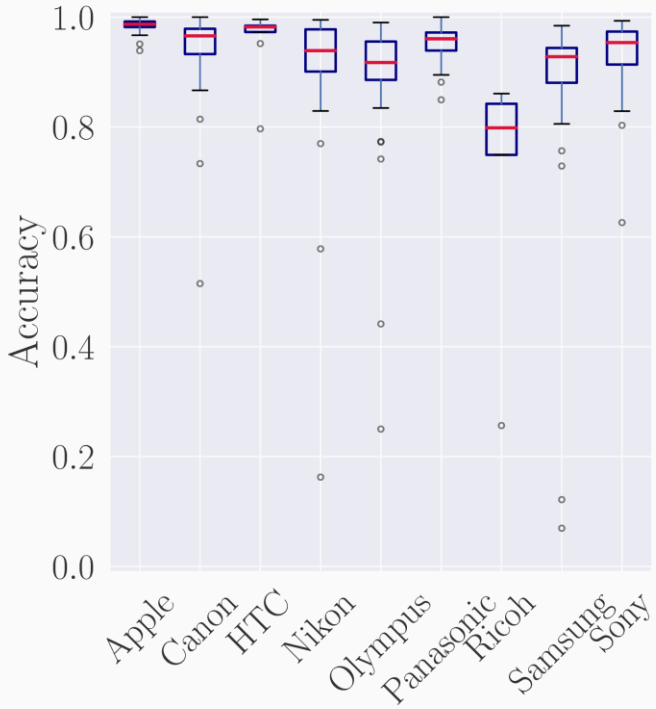
IFD histograms



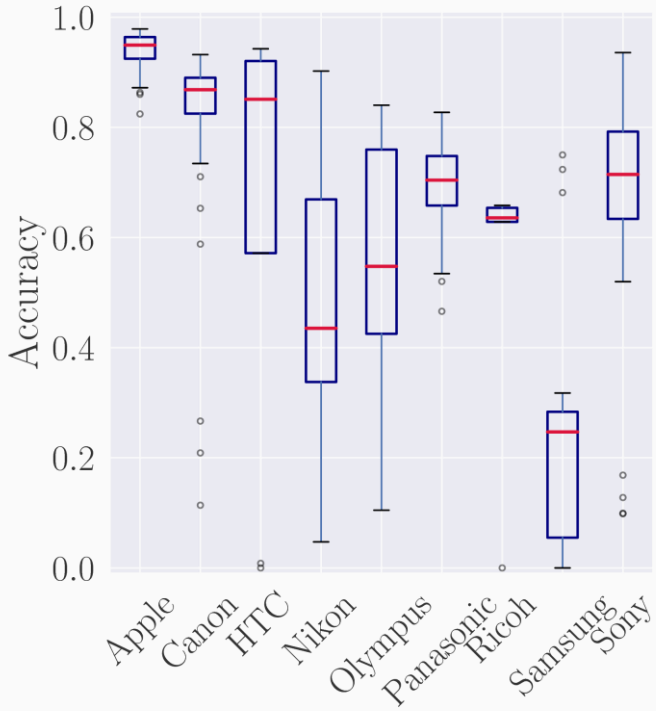
- Each plot shows accuracy of associating a unknown model with its make
- Each box within a plot represents one out of nine makes
- Best performance is achieved if IFDs are member of features (usually median > 90%)

Classification performance, considering different feature sets

IFD histograms



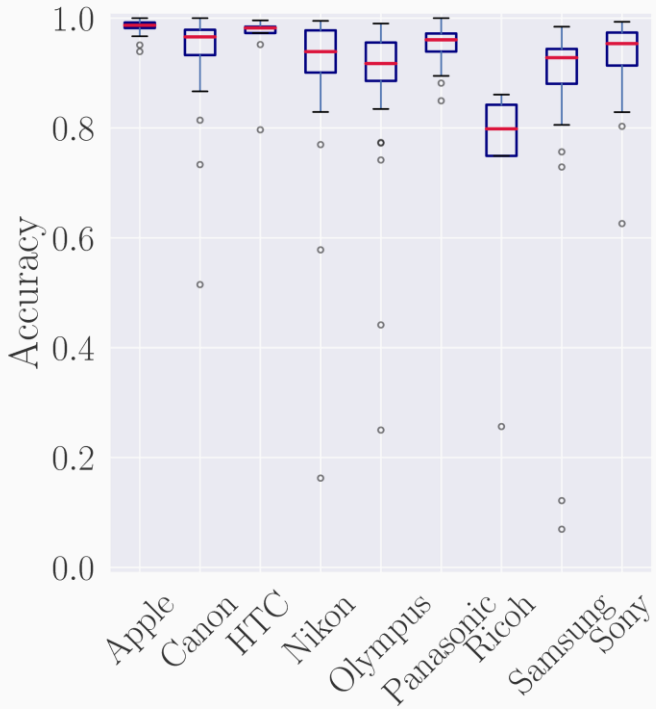
Quantization matrices



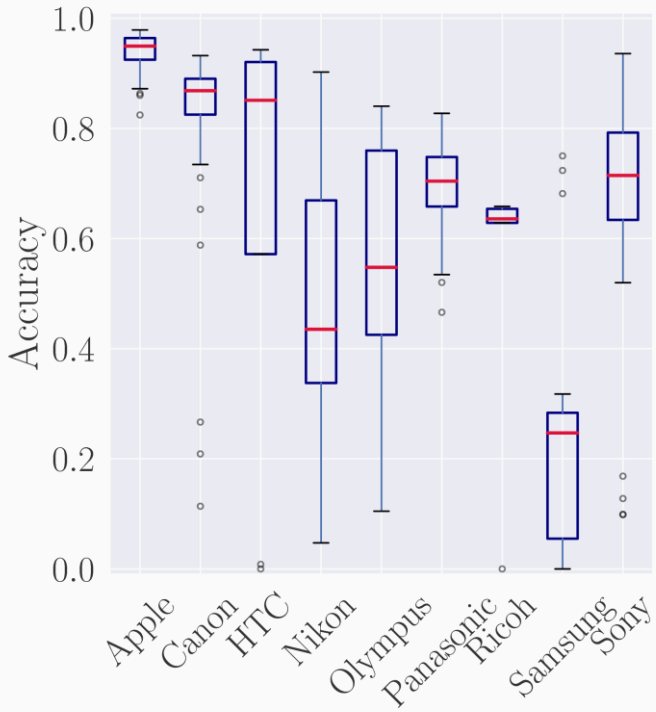
- Each plot shows accuracy of associating a unknown model with its make
- Each box within a plot represents one out of nine makes
- Best performance is achieved if IFDs are member of features (usually median > 90%)

Classification performance, considering different feature sets

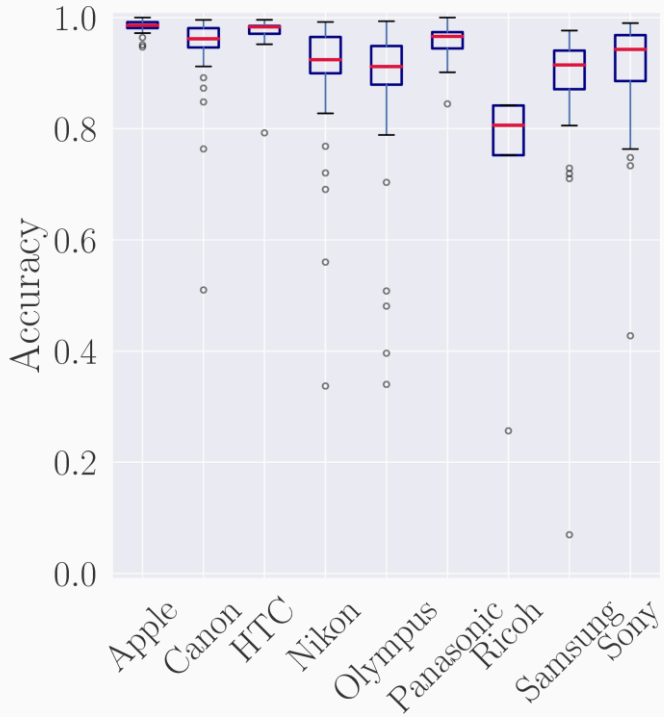
IFD histograms



Quantization matrices



Both feature sets



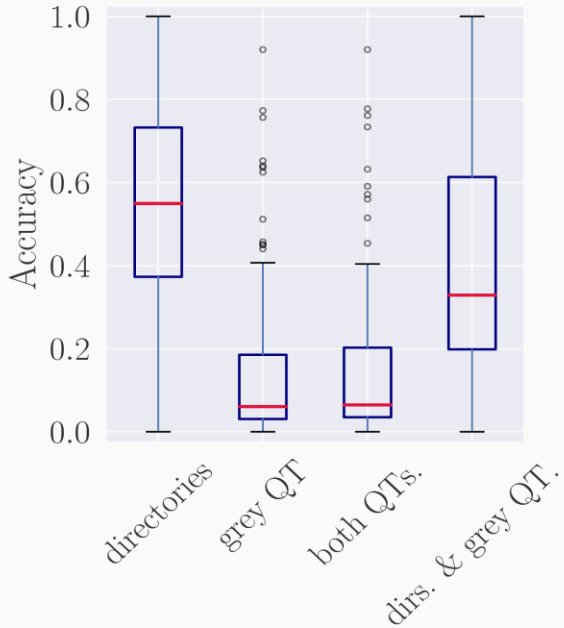
- Each plot shows accuracy of associating a unknown model with its make
- Each box within a plot represents one out of nine makes
- Best performance is achieved if IFDs are member of features (usually median > 90%)

Experiments

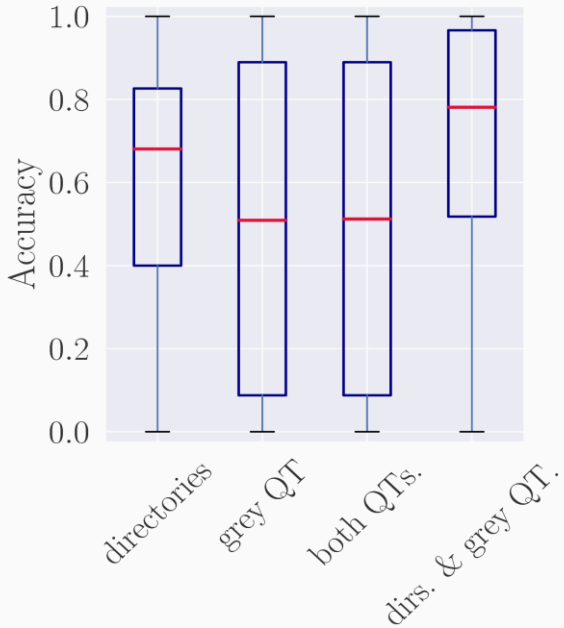
1. Predict make of images that stem from new, unknown, models
2. Predict make of images that were processed with additional software

Associating if software potentially edited header configuration

Processed by desktop software



Processed by mobile apps



- Processed images are more difficult to associate to a make
- Desktop software has a bigger impact
- We hypothesizes, this is due to limiting APIs the Phone platform offers

Summary

- We investigated source identification from file headers of JPEG images

Summary

- We investigated source identification from file headers of JPEG images
- Headers are extremely fast to read out and process

Summary

- We investigated source identification from file headers of JPEG images
- Headers are extremely fast to read out and process
- A common problem in source identification is the open-set problem

Summary

- We investigated source identification from file headers of JPEG images
- Headers are extremely fast to read out and process
- A common problem in source identification is the open-set problem
- Suggestion: apply hierarchical approach – Identify the make of a new model

Summary

- We investigated source identification from file headers of JPEG images
- Headers are extremely fast to read out and process
- A common problem in source identification is the open-set problem
- Suggestion: apply hierarchical approach – Identify the make of a new model
- This is successful due to intra-make similarities and inter-make variations

Summary

- We investigated source identification from file headers of JPEG images
- Headers are extremely fast to read out and process
- A common problem in source identification is the open-set problem
- Suggestion: apply hierarchical approach – Identify the make of a new model
- This is successful due to intra-make similarities and inter-make variations
- Especially histograms over the IFDs are excellent features for this task

Summary

- We investigated source identification from file headers of JPEG images
- Headers are extremely fast to read out and process
- A common problem in source identification is the open-set problem
- Suggestion: apply hierarchical approach – Identify the make of a new model
- This is successful due to intra-make similarities and inter-make variations
- Especially histograms over the IFDs are excellent features for this task
- The task is basically also solvable for processed images, however app based processing is less detrimental than desktop based processing

Summary

- We investigated source identification from file headers of JPEG images
- Headers are extremely fast to read out and process
- A common problem in source identification is the open-set problem
- Suggestion: apply hierarchical approach – Identify the make of a new model
- This is successful due to intra-make similarities and inter-make variations
- Especially histograms over the IFDs are excellent features for this task
- The task is basically also solvable for processed images, however app based processing is less detrimental than desktop based processing
- Further research, especially on the impact of additional software is required

Thank you