Approximate Algorithmic Image Matching to Reduce Online Storage Overhead of User Submitted Images

Braden D. Licastro

Allegheny College, USA

licastb@allegheny.edu

April 18, 2014

Photo Sharing Service

Definition

Photo sharing is the publishing or transfer of a user's digital photos online, thus enabling the user to share them with others [1].

Example Services

Remembering the Facts

- 500 Million images are shared daily as of May 2013 [2]
- Daily image shares are expected to double in 2014 [2]
- Approximately 20% of stored data is duplicate [3]
- Eliminating duplicates can save roughly $1.8 million annually at current sharing levels[1]

[1] Assuming 2013 averages of $.05 per gigabyte and 1MB image size[3].

Goals Reached

Website Creation

Created a flexible website framework that was able to imitate an image sharing service.

The Algorithm

Employed a series of checks and algorithms to find and eliminate duplicate and near-duplicate images at the time of upload.

Result Compilation

Website generates real time directory file count, directory size, and collects time taken per image upload.

Website Details

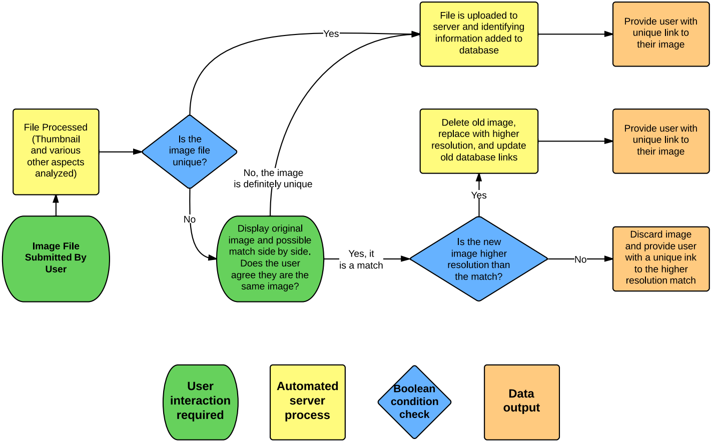
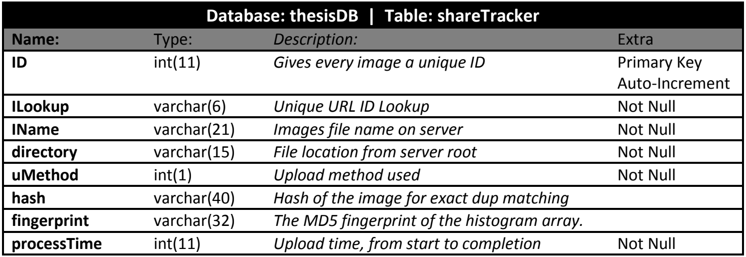


Image File Submitted By User
File Processed (Thumbnail and various other aspects analyzed)
Is the image file unique?
Yes
No
File is uploaded to server and identifying information added to database
Provide user with unique link to their image
No, the image is definitely unique
Display original image and possible match side by side. Does the user agree they are the same image?
Yes, it is a match
Is the new image higher resolution than the match?
Yes
Delete old image, replace with higher resolution, and update old database links
Provide user with unique link to their image
No
Discard image and provide user with a unique ink to the higher resolution match
User interaction required
Automated server process
Boolean condition check
Data output

Figure: Duplicate Identification Process

Website Details



Database: thesisDB | Table: shareTracker			
Name:	Type:	*Description:*	Extra
ID	int(11)	*Gives every image a unique ID*	Primary Key Auto-Increment
ILookup	varchar(6)	*Unique URL ID Lookup*	Not Null
IName	varchar(21)	*Images file name on server*	Not Null
directory	varchar(15)	*File location from server root*	Not Null
uMethod	int(1)	*Upload method used*	Not Null
hash	varchar(40)	*Hash of the image for exact dup matching*	
fingerprint	varchar(32)	*The MD5 fingerprint of the histogram array.*	
processTime	int(11)	*Upload time, from start to completion*	Not Null

Figure: Schema for File Uploads

Generation of Comparison Data

- Every upload is added to a baseline directory
- Only unique images are added to the duplicate-reduced directory
- Time taken for each upload is recorded in the database
- Real time directory statistics updated on every page load

Test Cases

All test cases are composed of half computer-generated, half photographic imagery.

I : Small Images $< 1MB$; No Duplicates

II : Small Images $< 1MB$; 20% Duplicates

III : Large Images $> 10MB$; No Duplicates

IV : Large Images $> 10MB$; 20% Duplicates

Performance Calculation

Using this equation, processing time, directory counts, and directory size were calculated.

$$\left(\frac{Base - Reduced}{Base}\right) * 100 = \%ImprovementOverBase$$

Figure: Percent efficiency over base case.

Processing Time

Additional processing time minimal given a data set containing unique and duplicate images.

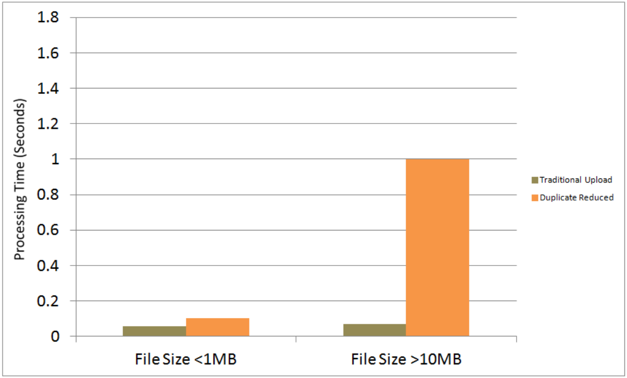


Figure: Processing time on data set with 20% duplicate images.

Processing Time

Worst case processing time is less than twice the previously observed.

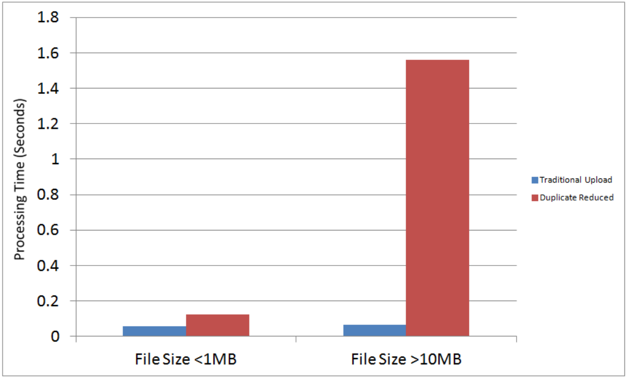


Figure: Processing time on data set with no duplicate images.

Storage Requirements

After running tests using an average of 20% duplicate image data...
Storage space was reduced approximately 12%.

Detection Rates

Identification patterns led to surprising conclusions:

- Correct identification of 82 of 100 small photographs.
- Identification of 91 of 100 large sized photographs.
- Surprisingly, only identified 33 of 100 computer generated graphics.

Identification Fault Explored

Low identification numbers were caused by the following:

- Large difference in image resolutions.
- 50/50 black and white distribution of coloring.
- Small thumbnail size.
- High detail repetitive patterns.

Dimension and Color Profiles

Resolution causing poor detection:

- Reducing image size causes data loss. Smaller images are impacted more significantly leading to poorer detection rates and false positives.

Color profiles and deep scans:

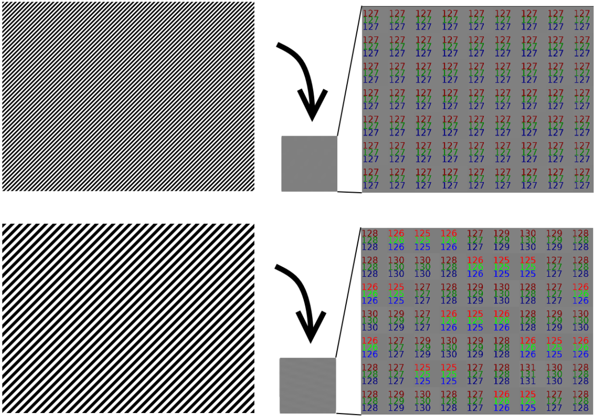
- Black and white images create frequent opportunity for identical color profiles.
- Identical color profiles require further investigation and re-sizing leading to above problem.

Further Details on Data Loss

Data loss through re-sizing:

- Reducing image size averages neighboring pixels into one destination pixel.
- Fine details will be lost in this process, possibly entirely.
- Thumbnail analysis may show false positives due to this.
- More prevalent with black and white repetitive patterns.

Data Loss Visualized



127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127
127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127
127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127
127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127
127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127
127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127
127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127	127 127 127

128 128 128	126 126 126	125 125 125	126 126 126	127 127 127	129 129 129	130 130 130	129 129 129	128 128 128
128 128 128	130 130 130	130 130 130	128 128 128	126 126 126	125 125 125	125 125 125	127 127 127	128 128 128
126 126 126	125 125 125	127 127 127	128 128 128	129 129 129	130 130 130	128 128 128	127 127 127	126 126 126
130 130 130	129 129 129	127 127 127	126 126 126	125 125 125	126 126 126	128 128 128	129 129 129	130 130 130
126 126 126	127 127 127	129 129 129	130 130 130	129 129 129	128 128 128	126 126 126	125 125 125	126 126 126
128 128 128	127 127 127	125 125 125	125 125 125	127 127 127	128 128 128	131 131 131	130 130 130	128 128 128
128 128 128	129 129 129	130 130 130	128 128 128	127 127 127	126 126 126	125 125 125	127 127 127	128 128 128

Figure: Visualized data loss after thumbnail creation of pattern.

Threats to Validity

Possible areas of concern needing addressed:

- Publicly accessible system introduces security vulnerabilities.
- Highly specific image modifications tested and may introduce bias.
- Undetected image corruption possible during upload.

Future Work

Recommended areas of future exploration include:

- Expansion of file support allowing for bmp, png, gif, etc.
- Further optimization of detection accuracy.
- Allow for other image manipulations.
- Additional testing, possibly live run for real world data.
- Determine frequency of hash collisions and impact on results.

Demonstration of Results

View the live website...

▸ System Login

References

Dictionary.com.
Definition of Photo Sharing.
www.dictionary.com, 2013.
[Online; accessed 20-November-2013].

All Things Digital.
Meeker: 500 Million Photos Shared Per Day and That's on Track to Double in 12 Months.
http://allthingsd.com/20130529/meeker-500-million-photos-shared-per-day-and-thats-on-track-to-double-in-12-months/, 2013.
[Online; accessed 20-November-2013].

NTP Software.
Survey Says Nearly Two-Thirds of Files on Primary Storage Are Stale.
www.ntpsoftware.com/pressroom/survey-says-nearly-two-thirds-files-primary-storage-are-stale, 2013.
[Online; accessed 31-October-2013].

Questions? Comments?