Image-to-image search with Pixcavator (PxSearch): a case study

27th February 2009

AssaySoft, Inc.
17195 Newhope Street, Suite 107
Fountain Valley, CA 92708 USA
www.assaysoft.com

Contents

Executive Summary

In this report we present the evaluation of the *Pixcavator Image Search* or *PxSearch*™ software. Presented with a query image *PxSearch* retrieves images from a collection of images based on their similarity to the query.

We received from GoDigital 83 original images and 15 folders corresponding to 15 categories of image transformations. Each folder contains a couple of original images modified by a given transformation. The goal is to determine if our software would be able to find the transformed image among the rest of the collection. A potential application of this software is a search for copyrighted images. The transformed images represent copies (legal or illegal) of the original image.

There are no standard evaluation criteria to measure the performance of image search software. We used the concepts of "recall" and "precision" widely used in the text search industry. Common web-based search engines like Google, Microsoft, and Yahoo also use similar criteria to measure the effectiveness of their software.

To see the effect of individual transformation, we build 15 separate databases which include the original 83 images and all the images of a transformation category. We selected the original image as a query image and evaluated if the software would be able to detect all the transformed images. We measured recall and precision for each of the 30 queries. The result of the analysis is provided in this report.

The average of values of recall and precision was computed and categorized according to the following criteria.

- Excellent = 80 – 100
- Satisfactory = 50 – 79
- Unsatisfactory = 0 – 49

Excellent

	Transformation	Average Score
1	Contrast	100
2	Writing	100
3	Hue	99
4	Rotation	97
5	Flip-Flop	96
6	HxW	96
7	Compression	96
8	De-Saturation	92
9	Skew	81

Satisfactory

	Transformation	Average Score
1	Saturation	65
2	Brightness	61
3	Close Cropping	54

Unsatisfactory

	Transformation	Average Score
1	FILTER	37
2	Off_Crop	18

It was not possible to produce results for the "Scale" transformation, because the collection does not contain the original image that was scaled.

These results can be independently verified with the supplied copy of *PxSearch*.

AssaySoft has another image search software called *iVision*™. This software is explained briefly in Appendix 2. *iVision* is built on the "Texture Feature Vectors" of the images. We expect that "Close Cropping" and "Off_Crop" categories will perform better with *iVision* because it is based on image tiling.

Overview of PxSearch (Pixcavator Image Search software)

PxSearch 1.2 is a program created for development and testing of our image search technology. We also developed commercial software called *cellAnalyst* which is an image analysis suite for scientific applications (see brochure in Appendix 1). The commercial versions of *PxSearch* will be built on the *cellAnalyst* platform.

The underlying image analysis methods of *cellAnalyst* and *PxSearch* are the same. Dr. Saveliev filed for a patent on the underlying methods in 2006 (patent application number 20070036434). *cellAnalyst* is available for a free download from our web site since Jan. 2008.

The purpose of the *PxSearch* software is to simply find images similar to a given query image. Since the concept of "similar image" is not very well defined, we focus our attention on finding modified version of the original image.

The system consists of the following modules:

- the collection of images that can be extended;
- the database containing "signatures" of images, images' origins, and other data;
- the image analysis unit (produces the signatures);
- the matching unit (matches the signatures);
- user interface (uploads an image, searches for similar images in the collection, displays the matches as a list).

Every image to be added to the collection is first converted to grayscale and then shrunk so that the larger dimension is 100 or 150. A version of *PxSearch* has the following option: several secondary versions of each image are created, analyzed, and added to the collection and their data is added to the database, total of 8:

- original
- rotation, 5 degrees
- rotation, 45 degrees
- Gaussian blur
- salt and pepper noise
- stretch, 5%
- shrink, 5%
- crop from all sides, 5%

This feature was not used in the report since GoDigital supplied modified copies.

The database entry for each image contains information about its origin, including:

- date and time;
- the filename of the original image;
- the way the image was produced from the original (shrinking, rotation, etc.);
- the signature of the image.

A *signature* is a sequence of integers which is the output of the image analysis. It is essentially the distribution of sizes of objects found in the image.

Suppose the signature of the two images are $\{A_n\}$ and $\{B_n\}$. Move along these sequences and compute the absolute value of the differences of n^{th} entries. The result is a distance formula as the "weighted 1-norm metric." It is expressed as:

$$D = \sum C_n \left| A_n - B_n \right|$$

This *distance* is intended to be a degree of separation between each reference image and each possible match in the test data set.

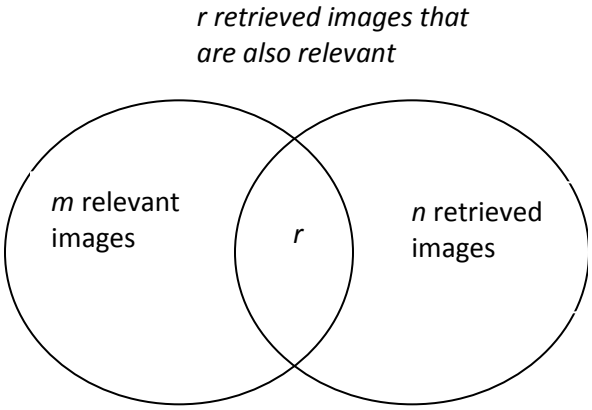
Informally, a search is deemed successful if most of the versions of the query image are at the top of the list.

Evaluation Criteria

Performance Measurement of Image Search

There are no standard evaluation criteria to measure the performance of image search applications. Therefore we have to choose a criterion to measure the performance of the image search that has its roots in the text retrieval industry. We use this approach initially and simply replace "documents" with "images".

Suppose for a given query there are "m" documents in the database. However, when the query is executed, only "n" documents are retrieved, out of those only "r" documents are relevant, i.e., good matches.



Then the following two measurements quantify the quality of the search:

Definition 1

Recall1 = *r* / *m* = Retrieved images that are also relevant / Total relevant images.

Precision1 = *r* / *n* = Retrieved images that are also relevant / Total retrieved images.

Ideally, the value of recall and precision should each be equal to 1. However, in the real world queries they behave inversely to each other. When the query is very broad, the recall is very high, but precision is low. When the query is very restrictive, the precision is very high and recall is very low.

Let's consider how these measurements are used for Google search. If we initiate a search as a very general query – give me all web sites where the word "computer" is written – we will get some relevant pages along with many which are not. In this case, recall is very high, but precision is low. If we search as a very specific query – give me all the web pages where the word "computer + PC + 2.2GHz CPU + 4 GB RAM" is written we will get few pages that are highly relevant, although some relevant ones may be missed. In this case precision is high, but recall is low.

In the search environment, the matches are simply ordered based on their distance from the query. We choose the distance of 70 as the *cut-off* or "threshold": all images within 70 from the query are declared matches and retrieved, the rest are not. The choice was made based on the examination of the search results in an attempt to include as many as possible of "good" matches and, at the same time, to exclude the matches that we judge "bad", even if they are modified versions of the original. Other values may be tried as well.

The main drawback of the above definition then is that it requires a cut-off to separate the retrieved images from the rest. Then, the evaluation results depend on this choice.

To mitigate this problem, we introduce a modified definition for recall and precision.

Definition 2

> *Recall2* = Total number of relevant images / Total count up to the last relevant image.
>
> *Precision2* = Total number of relevant images up to the first non-relevant image / Total number of relevant images.

Thus, we have 4 evaluation criteria:

- Recall1: Standard definition based on distance of 70;
- Precision1: Standard definition based on distance of 70;
- Recall2;
- Precision2.

We should point out that these are two very different definitions. If Recall1>Recall2, it does not mean that the quality is lower. These two numbers prove two independent evaluations of the quality of the search.

Is possible to set a threshold of what a 'correct' match is that doesn't also include incorrect matches? Clearly, we can't absolutely guarantee this unless we choose a very low cut-off and even then it may turn out to be too high if we increase the size of the image collection. In any case, Precision shows how close we are to this ideal.

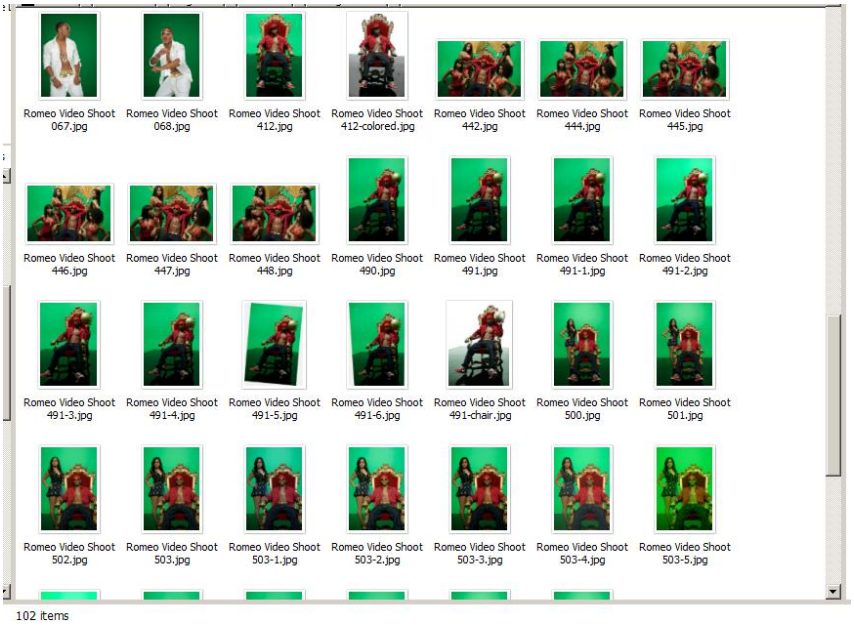
Throughout the report, the images marked yellow are the ones used for evaluation of the matching quality.

Example

The images in this example are taken from the initial collection provided by GoDigital in December 2008.

Ground Truth:

Suppose we have the following set of images.



Reference image	Description	Duplicate images	Count
798-REF	Group Dancing	1. 798 2. 798-2 3. 798-3 4. 798-4 5. 798-5 6. 798-6 7. 798-7 8. 798-8 9. 798-9 10. 798-10	12 + 1 (REF) = 13

		11. 798-11 12. 798.png	
46-REF	Man with white clothes	1. 46 2. 46-1 3. 46-2 4. 46-3 5. 46-4 6. 46-5 7. 46-6	7 + 1 (REF) = 8
503-REF	Man Chair + Woman Standing	1. 503 2. 503-1 3. 503-2 4. 503-3 5. 503-4 6. 503-5 7. 503-6 8. 503-7 9. 503-8	9 + 1 (REF) = 10
491-REF	Man Chair	1. 491 2. 491-1 3. 491-2 4. 491-3 5. 491-4 6. 491-5 7. 491-6 8. 491-Chair	8 + 1 (REF) = 9

Search results:

Query 1	Query 2	Query 3	Query 4
1. REF-798	1. REF-46	1. REF-503	1. REF-491
2. 798	2. 46	2. 503	2. 491
3. 798-2	3. 46-1	3. 503-7	3. 491-1
4. 798-4	4. 46-4	4. 503-8	4. 491-4
5. 798-3	5. 46-2	5. 503-1	5. 490
6. 798-5	6. 46-5	6. 503-3	6. 491-5
7. 798.png	7. 46-6	7. 503-5	7. 491-6
8. 798-8	8. 44	8. 502	8. 491-2
9. 798-7	9. 45	9. 503-4	9. 491-3
10. 798-6	10. 47	10. 504	10. 500
11. 796	11. 52	11. 503-2	11. 507
12. 794	12. 46-3	12. 503-6	12. 506
13. 798-10			13. 501
14. 795			14. 491.chair
15. 799			
16. 798-9			

17. 800			
18. 798-11			

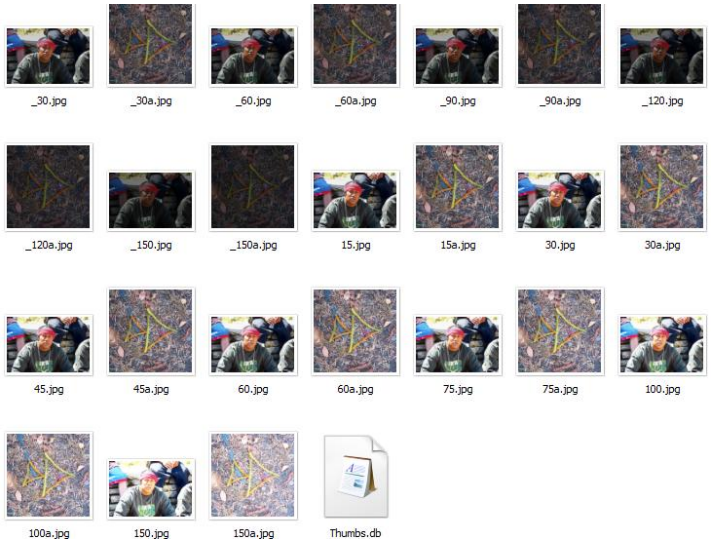
Recall2 = 13/18	Recall2 = 8/12	Recall2 = 10/12	Recall2 = 9/14
Precision2 = 10/13	Precision2 = 7/8	Precision2 = 7/10	Precision2 = 4/9

Analysis Data

Transformation: 01 Brightness

Ground Truth

Database: 83 original images + 24 duplicate images = 107 images



Original Image	Not in the database		Original Image	fish2.jpg
Description	Man with a red headband		Description	Dry leaves
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	_30.jpg		1	_30a.jpg
2	_60.jpg		2	_60a.jpg
3	_90.jpg		3	_90a.jpg
4	_120.jpg		4	_120a.jpg
5	_150.jpg		5	_150a.jpg
6	15.jpg		6	15a.jpg
7	30.jpg		7	30a.jpg
8	45.jpg		8	45a.jpg
9	60.jpg		9	60a.jpg
10	75.jpg		10	75a.jpg
11	100.jpg		11	100a.jpg
12	150.jpg		12	150a.jpg

Results

Query Image:

Rank	Distance	Image file name	

Query Image: fish2.jpg

Rank	Image file name	Distance
1	30a.jpg	18
2	15a.jpg	21
3	_30a.jpg	24
4	60a.jpg	31
5	45a.jpg	32
6	150a.jpg	43
7	_60a.jpg	49
8	*RomeoBlackBeat	50
9	100a.jpg	52
10	*fish5	58
11	75a.jpg	60
12	*Silkk2	65
13	_90a.jpg	75
23	_120a.jpg	93
35	_150a.jpg	101

File names designated with a "*" character have distance of less than 70 but are not duplicate of the original.

Recall1 =
Precision1 =

Recall1 = 9/12 = 75%
Precision1 = 9/12 = 75%

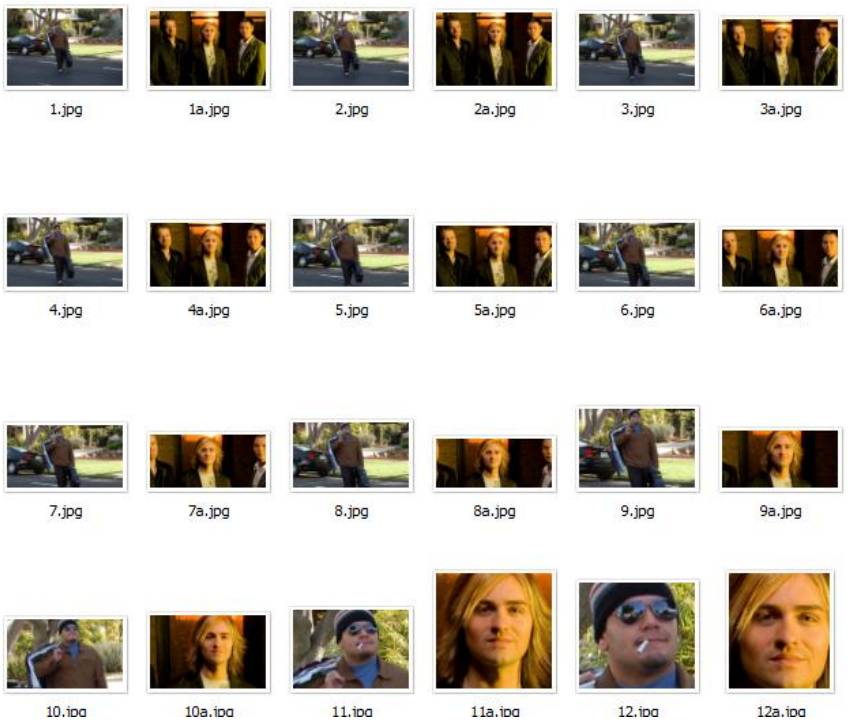
Recall2 =
Precision2 =

Recall2 = 12/35 = 34%
Precision2 = 7/12 = 58%

Transformation: 02 Close Cropping

Database: 83 original images + 24 duplicate images = 107 images

Ground Truth



Original Image	Picture 005.jpg		Original Image	01 STRIVE End Card.jpg
Description	Man with a suitcase		Description	3 men
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	1.jpg		1	1a.jpg

2	2.jpg		2	2a.jpg
3	3.jpg		3	3a.jpg
4	4.jpg		4	4a.jpg
5	5.jpg		5	5a.jpg
6	6.jpg		6	6a.jpg
7	7.jpg		7	7a.jpg
8	8.jpg		8	8a.jpg
9	9.jpg		9	9a.jpg
10	10.jpg		10	10a.jpg
11	11.jpg		11	11a.jpg
12	12.jpg		12	12a.jpg

Results

Query Image: Picture 005.jpg

Rank	Image file name	Distance
1	1.jpg	19
2	5.jpg	21
3	3.jpg	28
4	2.jpg	30
5	4.jpg	36
6	7.jpg	37
7	8.jpg	39
8	6.jpg	40
9	*Romeo Wallpaper	50
10	*DSC_0036	67
17	10.jpg	86
19	9.jpg	89
20	11.jpg	90
21	12.jpg	93

Query Image: 01 STRIVE End Card.jpg

Rank	Image file name	Distance
1	1a.jpg	35
2	4a.jpg	50
3	5a.jpg	56
4	3a.jpg	59
5	*Romeo_064_2	61
6	2a.jpg	61
7	*DSC_0855	66
16	6a.jpg	95
35	10a.jpg	124
41	7a.jpg	132
45	9a.jpg	135
49	8a.jpg	138
51	11a.jpg	140
92	12a.jpg	193

File names designated with a "*" character have distance of less than 70 but are not duplicate of the original.

Recall1 = 8/12 = 66%
Precision1 = 8/10 = 80%

Recall2 = 12/21 = 57%
Precision2 = 8/12 = 66%

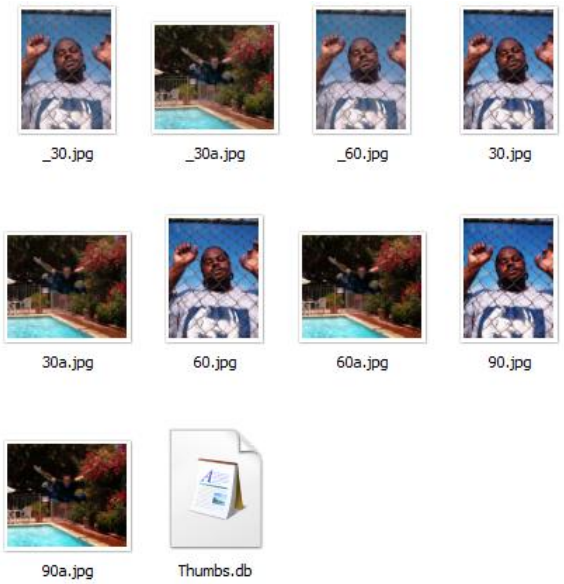
Recall1 = 5/12 = 42%
Precision1 = 5/7 = 71%

Recall2 = 12/92 = 13%
Precision2 = 4/12 = 33%

Transformation: 03 Contrast

Database: 83 original images + 9 duplicate images = 92 images

Ground Truth



Original Image	me.jpg		Original Image	Not in the database
Description	Man behind a fence		Description	Swimming pool
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	_30.jpg		1	_30a.jpg
2	_60.jpg		2	30a.jpg
3	30.jpg		3	60a.jpg
4	60.jpg		4	90a.jpg
5	90.jpg			

Results

Query Image: me.jpg

Rank	Image file name	Distance
1	30.jpg	35
2	60.jpg	46
3	_30.jpg	51

Query Image:

Rank	Image file name	Distance

4	_60.jpg	62				
5	90.jpg	67				

Recall1 = 5/5 = 100%
Precision1 = 5/5 = 100%

Recall2 = 5/5 = 100%
Precision2 = 5/5 = 100%

Recall1 =
Precision1 =

Recall2 =
Precision2 =

Transformation: 04 De-saturation

Database: 83 original images + 21 duplicate images = 104 images

Ground Truth


_10.jpg


_10a.jpg


_20.jpg


_20a.jpg


_30.jpg


_30a.jpg


_40.jpg


_40a.jpg


_50.jpg


_50a.jpg


_60.jpg


_60a.jpg


_70.jpg


_70a.jpg


_80.jpg


_80a.jpg


_90.jpg


_90a.jpg


_100.jpg


_100a.jpg


rome pics 735.jpg

Original Image	rome pics 735.jpg		Original Image	High School Romance.jpg
Description	Man with green shirt		Description	High School Romance
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	_10.jpg		1	_10a.jpg
2	_20.jpg		2	_20a.jpg
3	_30.jpg		3	_30a.jpg
4	_40.jpg		4	_40a.jpg

5	_50.jpg		5	_50a.jpg
6	_60.jpg		6	_60a.jpg
7	_70.jpg		7	_70a.jpg
8	_80.jpg		8	_80a.jpg
9	_90.jpg		9	_90a.jpg
10	_100.jpg		10	_100a.jpg
11	rome pics 735.jpg			

Results

Query Image: rome pics 735.jpg

Rank	Image file name	Distance
1	rome pics 735.jpg	0
2	_30.jpg	6
3	_10.jpg	9
4	_20.jpg	9
5	_50.jpg	12
6	_40.jpg	14
7	_60.jpg	18
8	_70.jpg	30
9	_80.jpg	31
10	_100.jpg	35
11	_90.jpg	36
12	*Romeo Part_2_14.jpg	69
13	*Rome pics 719.jpg	70

Query Image: High School Romance.jpg

Rank	Image file name	Distance
1	_10a.jpg	11
2	_20a.jpg	12
3	_30a.jpg	12
4	_40a.jpg	21
5	_50a.jpg	22
6	_60a.jpg	27
7	_70a.jpg	28
8	_80a.jpg	28
9	_90a.jpg	28
10	_100a.jpg	45
11	*p_romeo TEAM	47
12	*Header_top	58
13	*_10.jpg	62
14	*_20.jpg	63
15	*_30.jpg	64
16	*Rome pics 735	65
17	*_60.jpg	65
18	*Rome pics 735	65
19	*DSC 0813	66
20	*_40.jpg	68
21	*_50.jpg	69

File names designated with a "*" character have distance of less than 70 but are not duplicate of the original.

Recall1 = 11/11 = 100%
Precision1 = 11/13 = 85%

Recall2 = 11/11 = 100%
Precision2 = 11/11 = 100%

Recall1 = 10/10 = 100%
Precision1 = 10/21 = 47%

Recall2 = 10/10 = 100%
Precision2 = 10/10 = 100%

*Transformation: 05 FILTER_*

Database: 83 original images + 5 duplicate images = 88 images

Ground Truth


1.jpg


emboss.jpg


liq.jpg


sum.jpg


Thumbs.db


tiles.jpg

Original Image	rome pics 735.jpg		Original Image	
Description	Man with green shirt		Description	
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	1.jpg		1	
2	emboss.jpg		2	
3	liq.jpg		3	
4	sum.jpg		4	
5	tiles.jpg		5	
			6	
			7	
			8	

Results

Query Image: rome pics 735.jpg Query Image:

Rank	Image file name	Distance		Rank	Image file name	Distance
1	liq.jpg	15		1		
21	tiles.jpg	114		2		
38	sum.jpg	162		3		
43	1.jpg	184		4		
60	emboss.jpg	225		5		

				6		
				7		
				8		

Recall1 = 1/5 = 20%
Precision1 = 1/1 = 100%

Recall2 = 5/60 = 8%
Precision2 = 1/5 = 20%

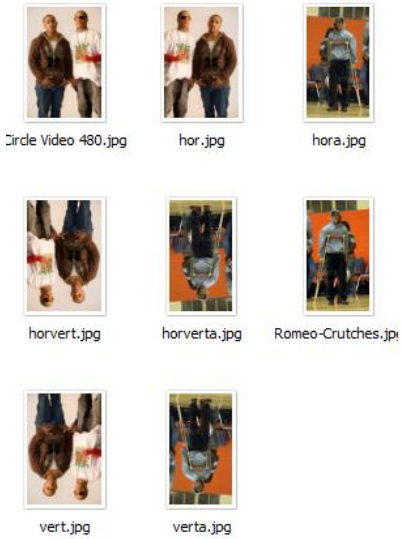
Recall1 =
Precision1 =

Recall2 =
Precision2 =

Transformation: 06 FlipFlop

Database: 83 original images + 8 duplicate images = 91 images

Ground Truth



Original Image	Circle Video 480.jpg		Original Image	Romeo Crutches.jpg
Description	2 men standing		Description	Man with crutches
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	Circle Video 480.jpg		1	hora.jpg
2	hor.jpg		2	horverta.jpg
3	horvert.jpg		3	Romeo Crutches.jpg
4	vert.jpg		4	verta.jpg

Results

Query Image: Circle Video 480.jpg

Rank	Image file name	Distance
1	Circle Video 480.jpg	0
2	hor.jpg	0
3	horvert.jpg	0
4	vert.jpg	0

Query Image: Romeo Crutches.jpg

Rank	Image file name	Distance
1	Romeo Crutches.jpg	0
2	hora.jpg	9
3	verta.jpg	9
4	horverta.jpg	21

5	*Romeo2 – gn4-sm2	63				
6	*DSC_0035	69				

File names designated with a "*" character have distance of less than 70 but are not duplicate of the original.

Recall1 = 4/4 = 100%
Precision1= 4/6 = 67%

Recall2 = 4/4 = 100%
Precision2= 4/4 = 100%

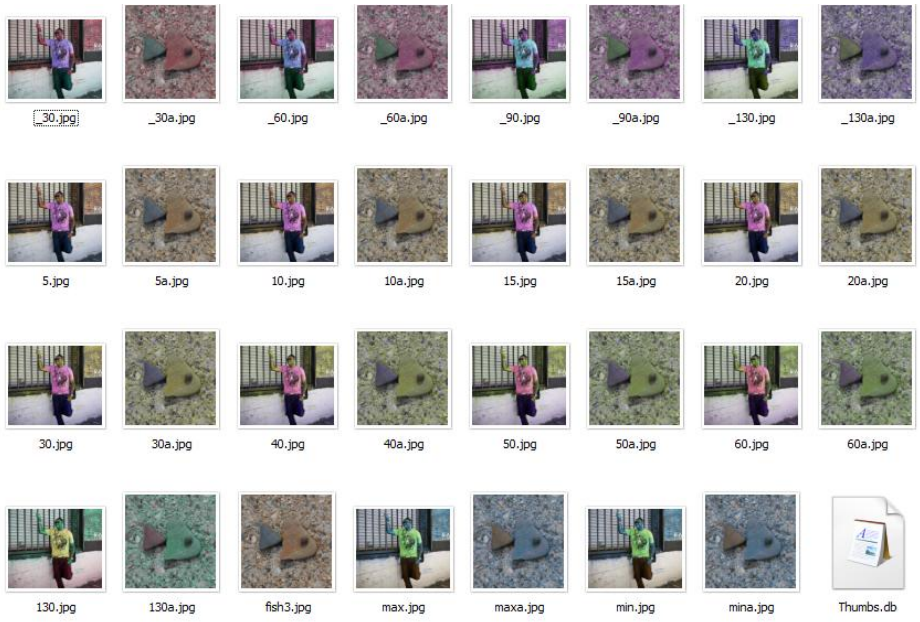
Recall1 = 4/4 = 100%
Precision1 = 4/4 = 100%

Recall2 = 4/4 = 100%
Precision2 = 4/4 = 100%

Transformation: 07 Hue

Ground Truth

Database: 83 original images + 31 duplicate images = 114 images



Original Image	Silkk2.jpg		Original Image	fish3.jpg
Description	Man standing on one leg		Description	3 stones
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	_30.jpg		1	_30a.jpg
2	_60.jpg		2	_60a.jpg
3	_90.jpg		3	_90a.jpg

4	_130.jpg		4	_130a.jpg
5	5.jpg		5	5a.jpg
6	10.jpg		6	10a.jpg
7	15.jpg		7	15a.jpg
8	20.jpg		8	20a.jpg
9	30.jpg		9	30a.jpg
10	40.jpg		10	40a.jpg
11	50.jpg		11	50a.jpg
12	60.jpg		12	60a.jpg
13	130.jpg		13	130a.jpg
14	Max.jpg		14	Fish3.jpg
15	Min.jpg		15	Maxa.jpg
			16	mina.jpg

Results

Query Image: Silkk2.jpg

Rank	Image file name	Distance
1	5.jpg	0
2	10.jpg	12
3	30.jpg	20
4	60.jpg	20
5	50.jpg	21
6	40.jpg	22
7	Max.jpg	23
8	Min.jpg	23
9	15.jpg	26
10	20.jpg	27
11	_30.jpg	29
12	_60.jpg	29
13	130.jpg	42
14	_130.jpg	48
15	_90.jpg	53
16	*RomeoXmas Paper	59

Query Image: fish3.jpg

Rank	Image file name	Distance
1	Fish3.jpg	0
2	5a.jpg	6
3	10a.jpg	10
4	130a.jpg	11
5	15a.jpg	16
6	60a.jpg	18
7	20a.jpg	21
8	30a.jpg	21
9	_90a.jpg	22
10	40a.jpg	24
11	50a.jpg	25
12	_60a.jpg	26
13	Maxa.jpg	27
14	Mina.jpg	27
15	_30a.jpg	28
16	_130a.jpg	34
17	*fish5	59

File names designated with a "*" character have distance of less than 70 but are not duplicate of the original.

Recall1 = 15/15 = 100%
Precision1 = 15/16 = 94%

Recall2 = 15/15 = 100%
Precision2 = 15/15 = 100%

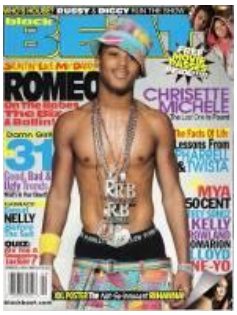
Recall1 = 16/16 = 100%
Precision1 = 16/17 = 94%

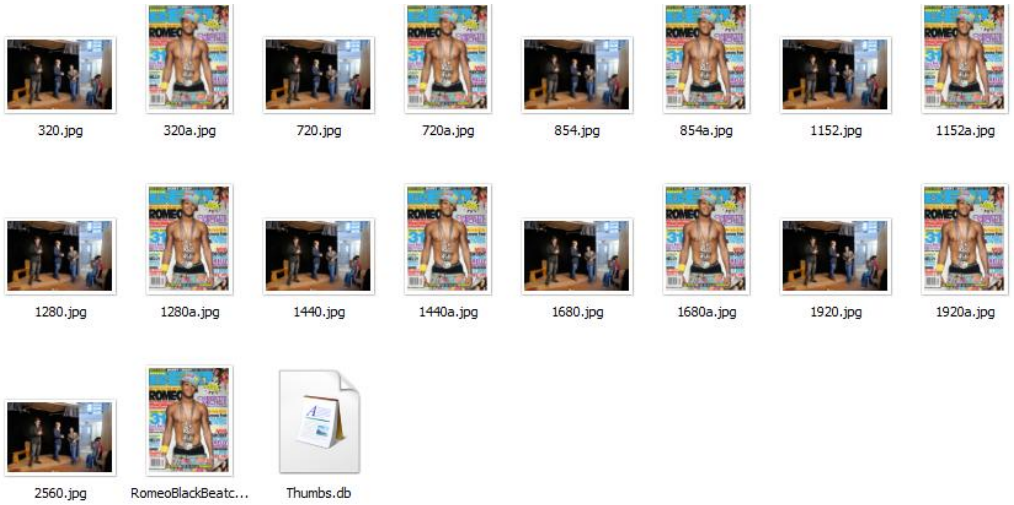
Recall2 = 16/16 = 100%
Precision2 = 16/16 = 100%

Transformation: 08 HxW

Database: 83 original images + 18 duplicate images = 101 images

Ground Truth





Original Image	DSC_0855.jpg		Original Image	Romeo Black Beat.jpg
Description	3 men and 1 woman		Description	Man with black belt
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	320.jpg		1	320a.jpg
2	720.jpg		2	720a.jpg
3	854.jpg		3	854a.jpg
4	1152.jpg		4	1152a.jpg
5	1280.jpg		5	1280a.jpg
6	1440.jpg		6	1440a.jpg
7	1680.jpg		7	1680a.jpg

8	1920.jpg		8	1920a.jpg
9	2560.jpg		9	Romeo Black beat.jpg

Results

Query Image: DSC_0855.jpg

Rank	Image file name	Distance
1	1920.jpg	5
2	1152.jpg	6
3	1440.jpg	6
4	1680.jpg	7
5	720.jpg	8
6	2560.jpg	8
7	1280.jpg	10
8	320.jpg	11
9	854.jpg	15

Query Image: Romeo Black Beat.jpg

Rank	Image file name	Distance
1	1920a.jpg	8
2	1440a.jpg	10
3	1280a.jpg	11
4	Romeo Black Beat.jpg	11
5	720a.jpg	16
6	1152a.jpg	16
7	1680a.jpg	16
8	854a.jpg	37
9	320a.jpg	43
10	*Fish2	57
11	*Silkk2	57
12	*BHHS BALL	67
13	*nRomeo Headshot	67
14	*Romeo lottery	70

File names designated with a "*" character have distance of less than 70 but are not duplicate of the original.

Recall1 = 9/9 = 100%
Precision1 = 9/9 = 100%

Recall2 = 9/9 = 100%
Precision2 = 9/9 = 100%

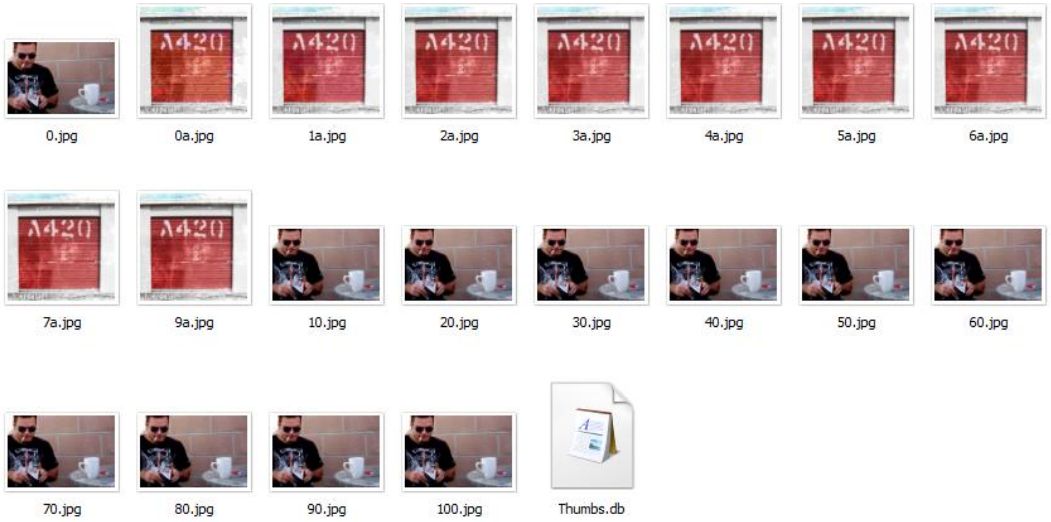
Recall1 = 9/9 = 100%
Precision1 = 9/14 = 64%

Recall2 = 9/9 = 100%
Precision2 = 9/9 = 100%

Transformation: 09 Compression

Database: 83 original images + 20 duplicate images = 103 images

Ground Truth



Original Image	Picture 054.jpg		Original Image	A420 Cover.jpg
Description	Man with a cigarette		Description	A420
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	0.jpg		1	0a.jpg
2	10.jpg		2	1a.jpg
3	20.jpg		3	2a.jpg
4	30.jpg		4	3a.jpg
5	40.jpg		5	4a.jpg

6	50.jpg		6	5a.jpg
7	60.jpg		7	6a.jpg
8	70.jpg		8	7a.jpg
9	80.jpg		9	9a.jpg
10	90.jpg			
11	100.jpg			

Results

Query Image: Picture 054.jpg

Rank	Image file name	Distance
1	40.jpg	0
2	50.jpg	0
3	70.jpg	0
4	90.jpg	0
5	100.jpg	0
6	10.jpg	1
7	20.jpg	1
8	60.jpg	1
9	80.jpg	1
10	0.jpg	2
11	30.jpg	2

Query Image: A420 Cover.jpg

Rank	Image file name	Distance
1	9a.jpg	1
2	3a.jpg	5
3	4a.jpg	5
4	5a.jpg	7
5	7a.jpg	8
6	6a.jpg	10
7	2a.jpg	11
8	1a.jpg	19
9	0a.jpg	33
10	*DSC_0033	61
11	*Rome pics 797	62
12	*Rome-pics-797	65
13	*rRomeoHeadShot	68
14	*Romeo-blue-bball	70

File names designated with a "*" character have distance of less than 70 but are not duplicate of the original.

Recall1 = 11/11 = 100%
Precision1 = 11/11 = 100%

Recall2 = 11/11 = 100%
Precision2 = 11/11 = 100%

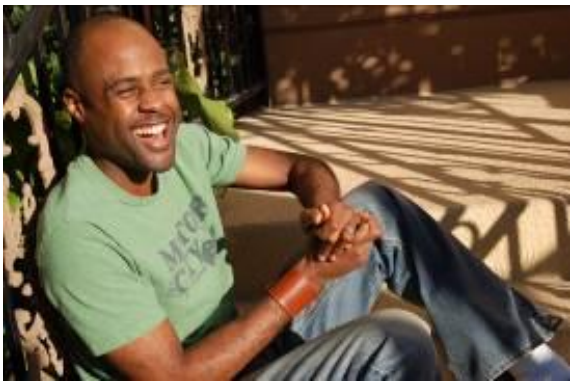
Recall1 = 9/9 = 100%
Precision1 = 9/14 = 64%

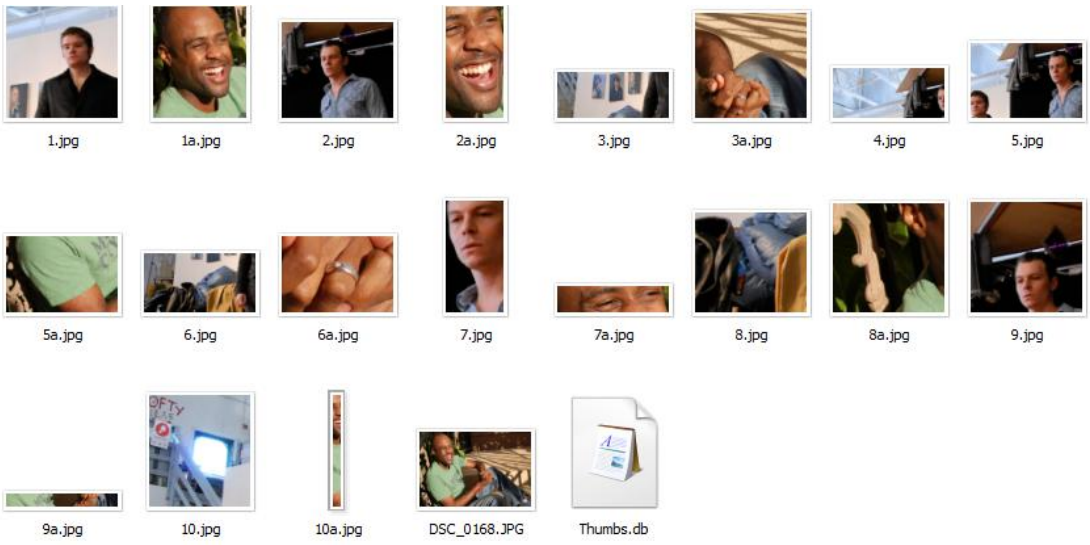
Recall2 = 9/9 = 100%
Precision2 = 9/9 = 100%

Transformation: 10 Off_Crop sub sec

Database: 83 original images + 20 duplicate images = 103 images

Ground Truth





Original Image	DSC_0824.jpg		Original Image	DSC_0168.jpg
Description	2 men		Description	Man sitting
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	1.jpg		1	1a.jpg
2	2.jpg		2	2a.jpg
3	3.jpg		3	3a.jpg
4	4.jpg		4	5a.jpg
5	5.jpg		5	6a.jpg
6	6.jpg		6	7a.jpg

7	7.jpg		7	8a.jpg
8	8.jpg		8	9a.jpg
9	9.jpg		9	10a.jpg
10	10.jpg		10	DSC_0168.jpg

Results

Query Image: DSC_0824.jpg

Rank	Image file name	Distance
24	2.jpg	158
30	7.jpg	166
37	6.jpg	174
38	9.jpg	174
46	1.jpg	177
51	5.jpg	180
52	8.jpg	180
60	4.jpg	187
96	3.jpg	268
99	10.jpg	291

Recall1 = 0/10 = 0%
Precision1 = 0/0 = 0%

Recall2 = 10/99 = 10%
Precision2 = 0/10 = 0%

Query Image: DSC_0168.jpg

Rank	Image file name	Distance
1	DSC_0168.jpg	0
3	3a.jpg	135
17	1a.jpg	179
28	8a.jpg	196
62	2a.jpg	248
75	5a.jpg	264
76	6a.jpg	271
95	7a.jpg	314
98	9a.jpg	327
101	10a.jpg	340

Recall1 = 1/10 = 10%
Precision1 = 1/1 = 100%

Recall2 = 10/101 = 10%
Precision2 = 1/10 = 10%

Transformation: 11 Rotation

Database: 83 original images + 13 duplicate images = 96 images

Ground Truth

Original Image	DSC_0029.jpg		Original Image	BHHS BALL 165.jpg
Description	Man with white shirt		Description	Basketball
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	a1.jpg		1	b1.jpg
2	a2.jpg		2	b2.jpg
3	a3.jpg		3	b3.jpg
4	a4.jpg			
5	a5.jpg			
6	a6.jpg			
7	a7.jpg			
8	a8.jpg			
9	a9.jpg			
10	a10.jpg			

Results

Query Image: DSC_0029.jpg

Rank	Image file name	Distance
1	a9.jpg	26
2	a2.jpg	27
3	a1.jpg	29
4	a3.jpg	34
5	a10.jpg	35
6	a8.jpg	38
7	a4.jpg	46
8	a5.jpg	52
9	a7.jpg	55
10	a6.jpg	63

Query Image: BHHS BALL 165.jpg

Rank	Image file name	Distance
1	b1.jpg	2
2	b2.jpg	2
3	b3.jpg	2
4	*RomeoXmasPaper	62

File names designated with a "*" character have distance of less than 70 but are not duplicate of the original.

Recall1 = 10/10 = 100%
Precision1 = 10/10 = 100%

Recall2 = 10/10 = 100%
Precision2 = 10/10 = 100%

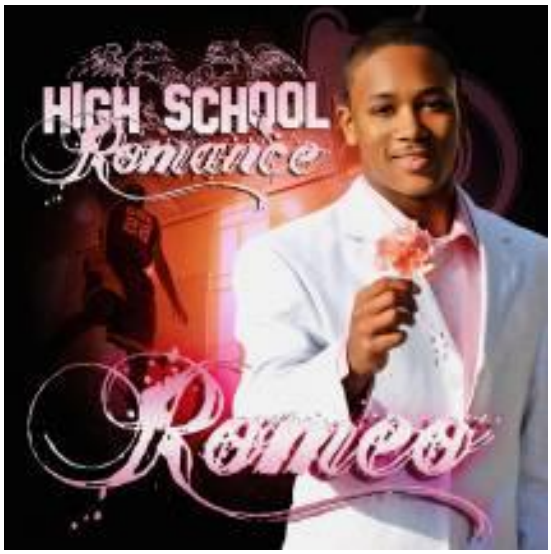
Recall1 = 3/3 = 100%
Precision1 = 3/4 = 75%

Recall2 = 3/3 = 100%
Precision2 = 3/3 = 100%

Transformation: 12 Saturation

Database: 83 original images + 20 duplicate images = 103 images

Ground Truth



Original Image	rome pics 735.jpg		Original Image	High School Romance.jpg
Description	Man with green shirt		Description	High School Romance
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	10.jpg		1	10a.jpg

2	20.jpg		2	20a.jpg
3	30.jpg		3	30a.jpg
4	40.jpg		4	40a.jpg
5	50.jpg		5	50a.jpg
6	60.jpg		6	60a.jpg
7	70.jpg		7	70a.jpg
8	80.jpg		8	80a.jpg
9	90.jpg		9	90a.jpg
10	100.jpg		10	100a.jpg

Results

Query Image: rome pics 735.jpg

Rank	Image file name	Distance
1	20.jpg	10
2	10.jpg	11
3	30.jpg	25
4	40.jpg	42
5	*Romeo_Part_2_14	69
6	*rome pics 719	70
18	60.jpg	97
32	50.jpg	114
50	70.jpg	158
54	100.jpg	166
56	80.jpg	169
57	90.jpg	169

Query Image: High School Romance.jpg

Rank	Image file name	Distance
1	50a.jpg	4
2	40a.jpg	9
3	20a.jpg	12
4	10a.jpg	13
5	80a.jpg	14
6	70a.jpg	17
7	30a.jpg	18
8	60a.jpg	20
9	100a.jpg	22
10	90a.jpg	23
11	*P_romeo TEAM	47
12	*Header Top	58
13	*20	60
14	*10	61
15	*40	61
16	*Rome pics	65
17	*DSC_0813	66
18	*30	69

File names designated with a "*" character have distance of less than 70 but are not duplicate of the original.

Recall1 = 4/10 = 40%
Precision1 = 4/6 = 67%

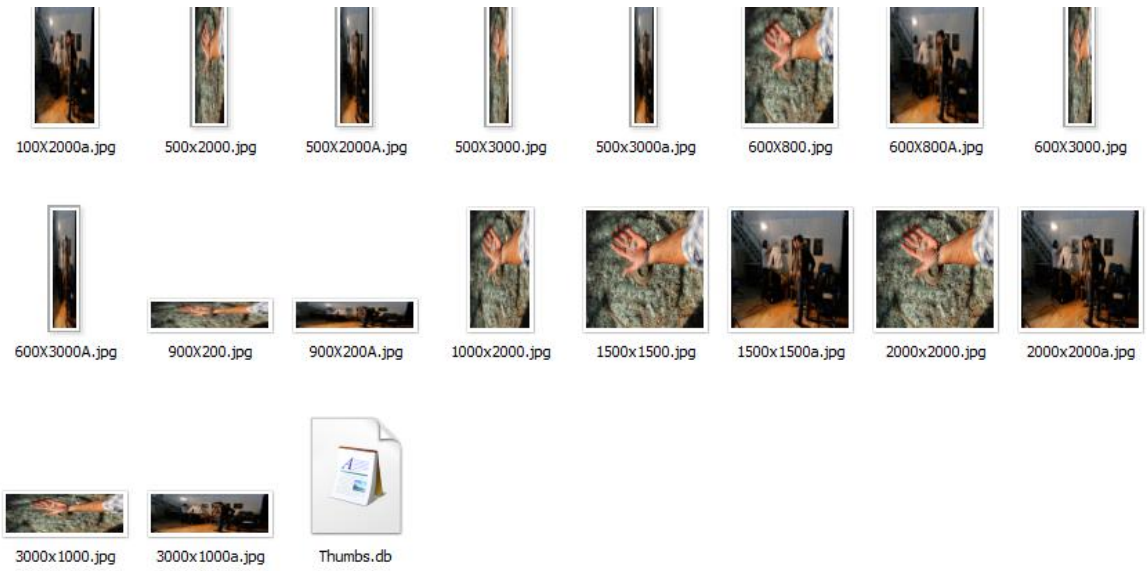
Recall2 = 10/57 = 18%
Precision2 = 4/10 = 40%

Recall1 = 10/10 = 100%
Precision1 = 10/18 = 56%

Recall2 = 10/10 = 100%
Precision2 = 10/10 = 100%

Transformation: 13 Scale

Ground Truth



Database: 83 original images + 18 duplicate images = 101 images

Original Image	Not in the database		Original Image	Not in the database
Description	Glass piece + hand		Description	3 men
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1			1	
2			2	
3			3	
4			4	
5			5	
6			6	
7			7	
8			8	
9			9	
10			10	

Results

Query Image: Query Image:

Rank	Image file name	Distance		Rank	Image file name	Distance

Recall1 =
Precision1 =

Recall2 =
Precision2 =

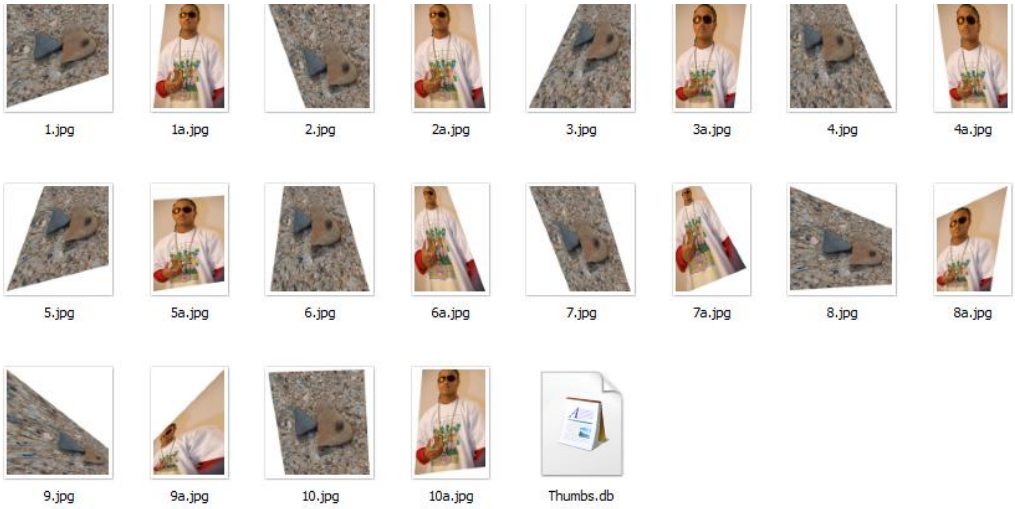
Recall1 =
Precision1 =

Recall2 =
Precision2 =

Transformation: 14 Skew

Database: 83 original images + 20 duplicate images = 103 images

Ground Truth



Original Image	fish3.jpg		Original Image	Circle Video 483.jpg
Description	3 stones		Description	Man with sun glasses
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	1.jpg		1	1a.jpg
2	2.jpg		2	2a.jpg
3	3.jpg		3	3a.jpg

4	4.jpg		4	4a.jpg
5	5.jpg		5	5a.jpg
6	6.jpg		6	6a.jpg
7	7.jpg		7	7a.jpg
8	8.jpg		8	8a.jpg
9	9.jpg		9	9a.jpg
10	10.jpg		10	10a.jpg

Results

Query Image: fish3.jpg

Rank	Image file name	Distance
1	10.jpg	29
2	4.jpg	36
3	6.jpg	43
4	2.jpg	44
5	3.jpg	45
6	8.jpg	50
7	1.jpg	56
8	*Fish5	59
9	5.jpg	60
10	9.jpg	67
11	7.jpg	75

Query Image: Circle Video 483.jpg

Rank	Image file name	Distance
1	10a.jpg	21
2	5a.jpg	30
3	3a.jpg	35
4	4a.jpg	50
5	6a.jpg	64
6	8a.jpg	67
7	7a.jpg	74
8	2a.jpg	83
9	1a.jpg	97
18	9a.jpg	133

File names designated with a "*" character have distance of less than 70 but are not duplicate of the original.

Recall1 = 9/10 = 90%
Precision1 = 9/10 = 90%

Recall2 = 10/11 = 90%
Precision2 = 7/10 = 70%

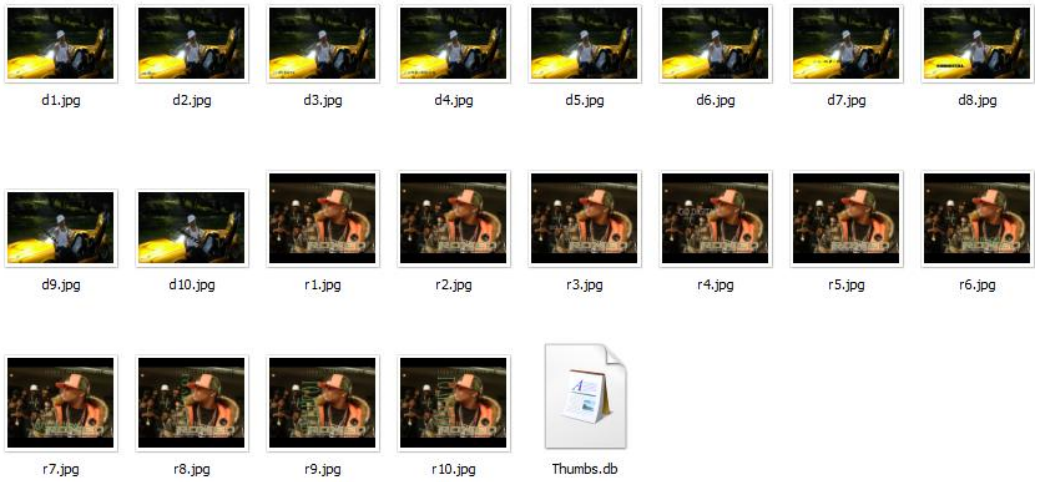
Recall1 =6/10 = 60%
Precision1 = 6/6 = 100%

Recall2 =10/18 = 56%
Precision2 = 9/10 = 90%

Transformation: 15 Writing

Database: 83 original images + 20 duplicate images = 103 images

Ground Truth



Original Image	DSC_0034.jpg		Original Image	Romeo Wall Paper 1280_1024-2.jpg
Description	Yellow Car		Description	Wall paper
Duplicate images	Duplicate Image file names		Duplicate images	Duplicate Image file names
1	d1.jpg		1	r1.jpg
2	d2.jpg		2	r2.jpg
3	d3.jpg		3	r3.jpg
4	d4.jpg		4	r4.jpg
5	d5.jpg		5	r5.jpg
6	d6.jpg		6	r6.jpg
7	d7.jpg		7	r7.jpg
8	d8.jpg		8	r8.jpg

9	d9.jpg		9	r9.jpg
10	d10.jpg		10	r10.jpg

Results

Query Image: DSC_0034.jpg

Rank	Image file name	Distance
1	d1.jpg	8
2	d5.jpg	9
3	d6.jpg	9
4	d9.jpg	9
5	d3.jpg	10
6	d4.jpg	10
7	d7.jpg	12
8	d2.jpg	18
9	d8.jpg	25
10	d10.jpg	36

Recall1 = 10/10 = 100%
Precision1= 10/10 = 100%

Recall2 = 10/10 = 100%
Precision2 = 10/10 = 100%

Query Image: Romeo Wall Paper 1280_1024-2.jpg

Rank	Image file name	Distance
1	r1.jpg	1
2	r6.jpg	1
3	r2.jpg	2
4	r5.jpg	2
5	r3.jpg	3
6	r7.jpg	3
7	r4.jpg	4
8	r8.jpg	5
9	r9.jpg	10
10	r10.jpg	11

Recall1 =10/10 = 100%
Precision1 = 10/10 = 100%

Recall2 =10/10 = 100%
Precision2 = 10/10 = 100%

Summary of Results

	Transformation	Recall1	Precision1	Recall2	Precision2	Average
1	Brightness					61
1	Brightness	75	75	34	58	
2	Close Cropping	66	80	57	66	54
2	Close Cropping	42	71	13	33	
3	Contrast	100	100	100	100	100
3	Contrast					
4	De-saturation	100	85	100	100	92
4	De-saturation	100	47	100	100	
5	FILTER	20	100	8	20	37
6	FlipFlop	100	67	100	100	96
6	FlipFlop	100	100	100	100	
7	Hue	100	94	100	100	99
7	Hue	100	94	100	100	
8	HxW	100	100	100	100	96
8	HxW	100	64	100	100	
9	Compression	100	100	100	100	96
9	Compression	100	64	100	100	
10	Off_Crop sub sec	0	0	10	0	18
10	Off_Crop sub sec	10	100	10	10	
11	Rotation	100	100	100	100	97
11	Rotation	100	75	100	100	
12	Saturation	40	67	18	40	65
12	Saturation	100	56	100	100	
13	Scale					
13	Scale					
14	Skew	90	90	90	70	81
14	Skew	60	100	56	90	
15	Writing	100	100	100	100	100
15	Writing	100	100	100	100	

	Average	80	81	76	79	79

Conclusions and Further Developments

Our main conclusion is that *PxSearch* can be successfully used for detecting mildly modified versions of images.

At the same time, we expect that further enhancement of the search algorithm will also improve the results. Also, modifications of the algorithm will allow the user to decide in what way (color, locations, etc) and to what degree the images are to be matched.

To overcome the poor performance of *PxSearch*'s algorithm in the cropping category, we may use our second image search software. It is called *iVision. And,* while *PxSearch* is based on the analysis of the topology of the image objects, *iVision* is built on the Texture Feature Vectors of the images. We therefore expect that "Close Cropping" and "Off_Crop" categories will perform better with *iVision* because it is based on image tiling.

Additional testing was conducted with GoDigital's image collection added to our own collection of about 5,000 images. The preliminary results suggest that our performance evaluation and the main conclusion above will hold.

Evaluation of the speed of the algorithm is not feasible with such a small image collections. It is important to note that in a distance-based search, the dependence of the time of search on the size of the collection is "logarithmic":

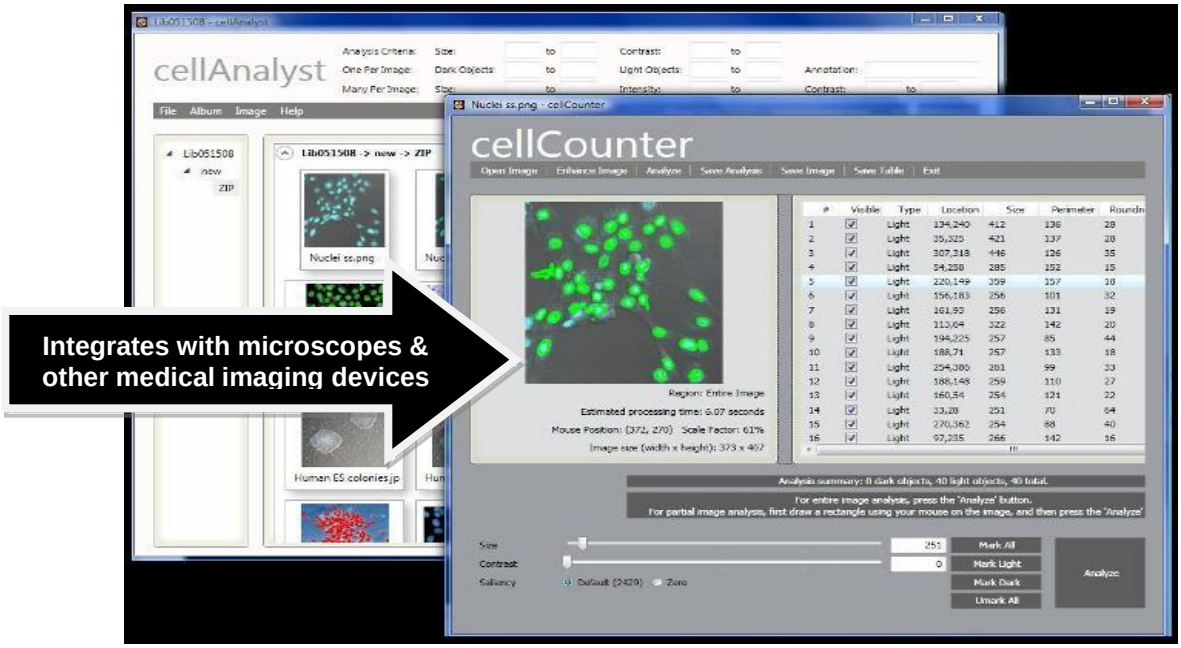
- if the collection grows 4X, the time grows 2X;
- if the collection grows 8X, the time grows 3X;
- if the collection grows 16X, the time grows 4X; etc.

Appendix 1: cellAnalyst Software Description

AssaySoft, Inc.
image mining software for high throughput microscopy

cellAnalyst ™

An advanced image-mining tool
in an exceptionally user-friendly package!



- **Images presented in familiar photo album format**
- **Cells detected, captured, & measured with the click of a mouse**
- **Cell images & characteristics displayed in table format**
- **Tables saved in searchable database**

Download your *free* copy of cellAnalyst 2.02
from www.AssaySoft.com
today!

cellAnalyst ™Is Ideal For

Bio-medical Imaging Facilities ● Pharmaceutical Research Labs ● High Content Analysis (HCA) for Drug Discovery

What *cellAnalyst* Does

- Analyzes images (captures and measures cells) and stores the output in a database.
- Summarizes this data across multiple images and collections.
- Queries the database and retrieves images with relevant cell information.
- Displays and manages images in the familiar photo album format.

For each image, cellAnalyst will produce the following output data:

- Image with cells' contours displayed.
- Spreadsheet with cells' locations and characteristics including area, perimeter, intensity, contrast, and many others.

The processing starts with an *automatic analysis* of the image that captures the complete data about the content of the image. An initial segmentation of the image is also provided.

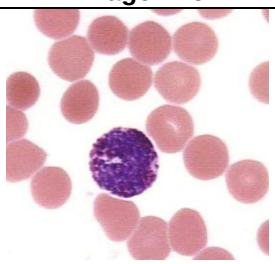
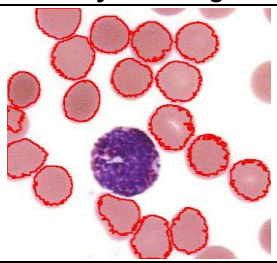
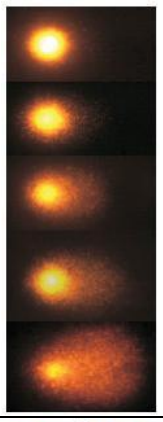
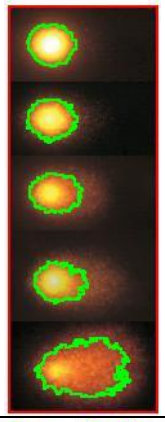
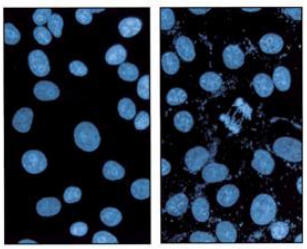
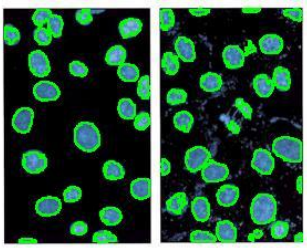
Next, in the *semi-automatic* mode, the user can interactively visualize multiple segmentations. By moving sliders corresponding to cell's characteristics, the user changes cells' contours and can choose the most appropriate segmentation. The output data is then updated instantly.

In the *manual* mode, the user can also exclude noise and irrelevant details from the analysis by simply clicking on them.

For every image in the user's collection, the complete set of image analysis data is saved as an entry in a database. This allows the user to search the entire collection for images with the desired content. For example, the user is able to retrieve all images that contain a given number of cells with given size, shape, and location. cellAnalyst also does the following:

- Visualizes biological standard 2D (TIFF, JPG, BMP, etc.), 3D, 4D, and 5D images.
- Provides tools to manipulate images by applying various filters and effects.
- Maintains version control of the images.

cellAnalyst Examples

Image File	Analyzed Image
	
	
	
	

These images were analyzed with cellAnalyst and the results are readily reproducible. Set the sliders to the correct values...a few mouse clicks...and cellAnalyst does the rest—*automatically!*

Appendix 2: *iVision* Software Description

iVision is an image database management system. It provides the ability to search images using content, metadata and annotation.

Album and Images

iVision allows the user to create a hierarchy of albums and images. The concept of albums and images is identical to the concept of folder and files of a file system. Just like a folder may contain other folders and/or files, an album may contain other albums and/or images. The reason we choose to call these data structures as albums and images because these are more appropriate words in the context of images.

When an image file is copied into the *iVision* system, the software maintains a copy of that image in its archives. If the user deletes that image file from the file system after copying in the iVision system, a copy is maintained in *iVision* and its functionality is not affected.

Content

The content search strategy starts by first computing the feature vectors of all the images stored in the database. An image is divided into tiles and a feature vector of every tile is computed. The tile size is dependent on the size of the image. The features vectors are stored in a database. The process of computing the image feature vectors is called ingestion.

During the search process the user identifies a tile by drawing a rectangle on the query image. Then the user specifies the feature descriptor to be used and the number of images needs to be retrieved from the database. The system computes the feature vector of the query tile and then computes the distance between the query tile and all the tiles of all the images in the database. Those images which contain tiles closest to the query image tile are retrieved from the database.

Annotations

Annotation can be added to any image in the album. Annotations are created by drawing a rectangle on the image and typing text that describes that portion of the image.

A dictionary is also maintained in the database to store domain specific words. For example, a biological database may choose to store words like “cell”, “tissue” etc. in a dictionary. The main advantage of maintaining a dictionary is that spelling errors can be avoided in annotation text entry. Also dictionary provides a common vocabulary for the group of people using *iVision* system and sharing their images.

After annotations have been entered, they can be edited and deleted. Annotations are searched using free text.

Metadata

Metadata is data about the images. Metadata can be loaded in the database from an Excel spreadsheet or an XML file. Once the metadata is entered, it can be modified and deleted from the database. Images can be searched using metadata.

iVision also provides tools to edit the database structure that contains the metadata. This means that columns in that table can be added, modified and deleted.

Digital Image Processing

Besides image archival and retrieval *iVision* also provides image processing facility. The following features are provided:

- Change Brightness / Contrast
- Change Color values
- Gamma control
- Equalize

- Filters
    - Edge enhancement
    - Sharpness
    - Blur

- Plot Histograms for Red, Green, Blue and Gray colors

Appendix 3: Biographies of AssaySoft's Dr. Ash Pahwa and Dr. Peter Saveliev

Ash Pahwa (Co-founder and CEO)

Dr. Ash Pahwa's expertise includes image processing, database management, digital video and data storage technologies. Dr. Pahwa has over 25 years of experience in the industry. He has held positions with General Electric, AT&T Bell Laboratories, Xerox Corporation, and Oracle. At Mayachitra, Inc. he developed an image database management system for storage and retrieval of biomedical images based on metadata, annotation and content. Dr. Pahwa is also the founder of DV Studio, Inc., which introduced successful products for CD-Recording (CDR), MPEG encoding, and videotape-to-DVD archiving at a time when the cost for entry into these technologies was far greater than it is today. DV Studio's CD-Gen was bundled with CD-Recorders by Philips, Yamaha and Sony. By establishing a level of usability and functionality for CDR, CD-Gen helped catapult it into the mainstream. Dr. Pahwa's book, "CD-Recordable Bible" (Pemberton Press, 1994), was translated into Japanese and German. He is listed in "Who's Who in the Frontiers of Science and Technology". He holds a Ph.D. in Computer Science from the Illinois Institute of Technology in Chicago.

Peter Saveliev (Co-founder and R&D Chief)

Dr. Peter Saveliev is a professor of mathematics at Marshall University, Huntington WV. Dr. Saveliev has conducted research on algebraic topology and a number of other fields. Prior to earning his Ph.D. from the University of Illinois at Urbana-Champaign, he worked as an applied mathematician and software engineer at large R&D institute in Moscow, Russia. There he developed numerical algorithms of image matching for missile guidance systems. After that he worked at a start-up company on an automated fingerprint identification system. Currently he develops applications of algebraic topology in computer vision and digital image analysis. He also runs wiki: Computer Vision Primer and blog: Computer Vision for Dummies.