Quality Measures for Interactive Image Retrieval with a Performance Evaluation of Two 3x3 Texel-Based Methods

D.P.Huijsmans*, M.S.Lew*, D.Denteneer**
*Computer Science Department, University of Leiden
P.O.Box 9512, 2300 RA Leiden
**Philips Research Lab, Eindhoven
The Netherlands
huijsman@cs.leidenuniv.nl,mlew@wi.leidenuniv.nl,dentenee@natlab.research.philips.com

Abstract. The aim of the Leiden Imaging and Multi-media Group in collaboration with Philips is to develop and evaluate content-based indexing and interactive retrieval methods for large photo collections and to integrate them with annotation based methods. Ground-truth is provided by copy pairs in the Leiden Portrait Database, a database of scanned-in images of 19th-century Dutch studio portraits ("Cartes de Visite").

Our highly effective projection vector indexing method is compared with Virage Datablade and two binary texel (3x3 B/W patterns) statistic feature vectors: a reported well-performing Local Binary Pattern and our 2D binary gradient pixel Trigram.

Evaluation criteria, based upon the number of copies found back within the visible top $\lfloor {}^2\log n \rfloor$ ranks, were defined for interactive internet image retrieval and applied to the ranking results for the test-set of 50 copy and 12 similar pairs embedded in 5570 portraits and studio logo images. Our evaluation shows that the projection method beats the binary texel based methods and Virage Datablade; the Trigram method performs better than the LBP method. Feature vector length reduction by grouping texel patterns in symmetry groups reduces the strenght of the Trigram method, whereas a KLT transform can be used to reduce the length of each feature vector by an order of magnitude without affecting the performance.

At "http://ind156b.wi.leidenuniv.nl:2000/" our visual search demo can be tried.

1 Query By Pictorial Example (QBPE)

As we mentioned earlier [Huijsmans96a], a straightforward generalization of text-retrieval methodologies for image retrieval is hampered by the fact that image annotation is incomplete and subjective (if not missing altogether) and that image content is always accompanied by lots of noise, preventing exact matches. Because of the often lacking right description text searching techniques are not

enough for image databases; forms of browsing and content-based retrieval have to be offered as well. We call this the **Visual Search ABC**: A for Annotation or Attributes, B for Browsing and C for Content-based retrieval. We develop tools for effective visual inspection: this article addresses the performance of content-based query by example techniques.

2 The similarity matching methods (feature vectors)

The similarity matching methods used in this article are based upon projection vectors [Huijsmans96a], 2D binary Pixel Trigrams in thresholded gradient images [Huijsmans96b] and Local Binary Patterns in intensity images [Ojala96]. The projection method uses the average row- and column values (line integrals) as a feature vector; an image of size $n \cdot m$ pixels gives rise to a horizontal and vertical projection vector of length $n+m$. Because [Ojala96] described a highly successful Trigram like feature vector called LBP (Local Binary Pattern) based on a method advocated by [Wang90] we used this one as a competitive local texture method for our own method. The Trigram and LBP methods use feature vectors based on the frequencies of binary 3x3 texel patterns in gradient and intensity space after thresholding with a noise level (Trigrams) or the central intensity value (LBP). In figure 1 the construction of a specific pattern index for one of the 512 possible Trigram patterns and one of the 256 possible LBP patterns is shown. For every 3x3 neighborhood in the image the pattern index-number is determined and used to add 1 to the specific pattern counter in the 512 or 256 element feature vector. Because in the LBP method the central 3x3 pixel does not contribute to the patternindex, the full feature vector length of LBP is twice as short as the full 3x3 based Trigram feature vector. After the formation of the normalised feature vector (Trigram/LBP pattern counters) the similarity matching is carried out by summing the absolute differences of the corresponding feature vector elements for every pair of images (L1-norm) and by sorting the obtained distances per row or column in the distance matrix by magnitude.

3 Optimizing feature vector length

Once an algorithm for extracting a feature vector is proposed that turns out to perform well, the question arises whether all of its elements are as important for this performance or whether about the same performance can be obtained by using a specific subset or weighted combination.

In [Lew96] a KLT (Karhunen-Loeve Transform) or Fisher LDA (Linear Discriminant Analysis) transformation turned out to perform as well as the projections method when only about 25 of the most variant coefficients were used. For each similarity matching method the length of the feature vector can thus be shortened by reducing the feature vector length to its most important KLT or LDA part.

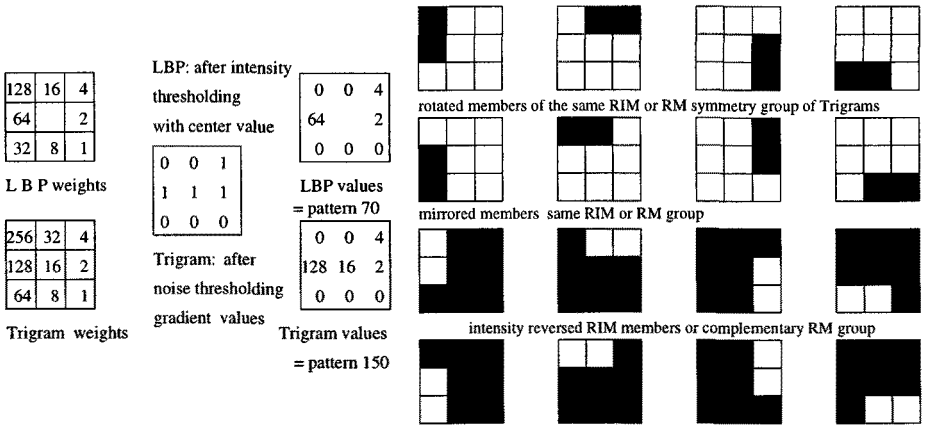

Fig. 1. (left) formation of the LBP and Trigram pattern index from a 3x3 binary texel (right) Trigram pattern symmetry group: 1 RIM, 2 RM groups

In [Huijsmans96b] weighing schemes were described to reduce the length of the feature vector: an effective weighing scheme was a band-pass filter that suppressed all but 75 pattern counts during the comparison.

Another way to reduce the number of LBP or Trigram patterns is based on the formation of rotation-, mirroring- and intensity invariant groups. For a review of invariant pattern recognition see [Wood96]. Figure 1 also shows one of the RIM (Rotation,Intensity,Mirroring symmetric) groups is, which is comprised of two RM (Rotation and Mirroring) groups that have complementary intensity values. By using the RIM groups of Trigram counts the full length feature vector reduces to 51. By forming RIM groups the number of different Trigrams is reduced to 51; when we confine the formation to RM groups twice as many, 102, Trigram pattern groups remain. The effect of symmetry grouping on the performance of the shortened Trigram feature matching was evaluated for a test-set of copy pairs and similars.

4 Performance evaluation criteria for interactive QBPE

The problem in evaluating image indexing and retrieval performance for similarity matches lies in the non-exactness of the matches and the lack of clearly defined ground-truth for similarity. To evaluate the performance of content-based similarity matching one would like:

- to have a clear definition of similarity.
- a noise threshold to distinguish similars from dissimilars.

Because similarity like beauty is in the eye of the beholder, anything but copy retrieval lacks an objective ground-truth. Although for copy retrieval noise thresholds can be set, this threshold depends upon the class of images; in our case we have two distinct classes (the front- and back-sides of studio portraits) and

the threshold level between these two classes for copy retrieval differs by about a factor of 3. Similarity matching for images will therefore never be more than a comparative ranking with the eye of the beholder (the user-interface) to set the threshold. Instead of trying to circumvent this fundamental threshold selection problem we suggest to incorporate the need for a user-controlled threshold setting in the retrieval procedure (human in the loop).

To prevent the need for scolling through similarity ranking results as much as possible, internet image retrieval should develop methods that usually deliver the right images among those shown on the first page of returned image thumbnails. No more than about 30 thumbnails can be shown on one html page to keep enough detail and preferably not more than about 10 per view for rapid evaluation and speed.

4.1 Visible Fraction F_v, visible Position P_v and retrieval Quality Q_r

Because we wanted the performance measures and the visible ranks to be some function of the database size without growing out of hand for very big databases we choose a logarithmical visible window size of length $L = \lfloor {}^2\log n \rfloor$, with n=database size. This means that for our present size of 5570 files the number of visible ranks will be 12; when at full size (100,000) the 17 top ranks will be displayed. With this logarithmical window even for databases of 1 to 100 million images no more than 20 to 27 images will have to be shown.

The most important performance measure for a user-interface driven retrieval system is how many of the test-pairs T have counterparts that appear in the top $L = \lfloor {}^2\log n \rfloor$ ranks: the number of these visible test-pairs are called (T_v). So

$$F_v = T_v/T$$

and is normalized to lie within [0,1]. We call this fraction F_v (**visible Fraction**) and it is considered to be the most important measure since it indicates how often copies can be found in the first view shown after a search has been specified.

A second performance measure is the average rank R_v for the visible test-pairs; from this average rank we can derive a normalized **visible Position** measure P_v which is defined as the ranking error divided by the length L of the display window. So

$$P_v = (L - R_v)/(L - 1)$$

with $L = \lfloor {}^2\log n \rfloor$. P_v lies within [0,1]; 0 when $R_v = L$ (all test-pairs just visible) and is 1 when $R_v = 1$ (all visible test-pairs on top); $P_v = 0$ when F_v=0.

Finally a combined **retrieval Quality** Q_r is defined by averaging the visible fraction and visible position measures:

$$Q_r = (F_v + P_v)/2$$

Q_r is also normalized to lie between [0,1] or [0, 100 %].

5 Image retrieval performance results

5.1 Copy pairs with projection, Virage Datablade and texels

A test-set of $T = 50$ copy pairs provided the ground-truth for the performance evaluation of the matching methods. Non-trivial copy pairs can be found in 19th century studio portraits because those portraits were usually delivered by the dozen; copies may have become quite different over the last century due to differences in exposure (bleeching), handling (dirt), trimming and staining. Each of the 50 copy pairs originate from a common negative and were delivered together. Parameters of the similarity matching methods are the different:

- pixel value domains: intensity-, gradient- or thresholded-gradient values.
- resolutions employed:75 ,37.5 or 18 dpi version of the scanned-in portraits (scanned-in at 300 dpi average size 770x1275).
- features:projections (horizontal and vertical), Local Binary Pattern statistics or 2D binary pixel Trigram statistics.
- feature vector length reduction schemes: KLT or band-pass.
- feature vector element weighing schemes: equal weights, anti-linear weights.
- distance metrics: L1- or L2-norm.
- symmetry groupings of trigrams: none, rotation and mirroring (RM), rotation and mirroring and intensity (RIM)

The distance matrices of 5570x5570 image comparisons each were calculated for about 30 methods. For each image the top 60 ranks for each method are stored (precalculated ranking results) in a Postgres6.0 database together with studio annotation. An initial evaluation of the methods tried led us to the three main methods compared in this section, which were those giving the best results overall. We wanted to show the performance of projections, LBP and Trigrams as a function of resolution and picked the best combination of other parameters.

The projections gave the best results when applied to gradient or binarized gradient images; the Trigram method was applied to pattern fractions weighted with a band-pass filter restricting the number of non-zero weights to about 75; the LBP method performed best at full length with equal weights. By applying a KLT transform, the feature vector elements can be ordered on variance and comparisons showed that the same matching performance could still be obtained when using only 10 % of the most important feature elements.

In all cases, projections, Trigrams and LBP, the L1-norm gave a better performance for similarity matching within these graylevel images than the L2-norm. The applied metric in Virage Datablade comparisons as well as the length of their feature vector is unknown.

In table 1 the resulting measures for the 3 methods at 3 resolution levels are given:

The projection method clearly performs best and almost optimal (both the visible fraction and the visible position is close to perfect); the Trigram method with 75 features intact outperforms the LBP method with 256 features. Using the KLT variance ordered feature elements both Trigram and LBP performance

Table 1. Copy retrieval performance of projection and texel methods in top $\lfloor^2\log n\rfloor$ ranks: $T = 50$ copy pairs embedded in $n = 5570$ files

Resolution in dpi	75	37	18	75	37	18	75	37	18	37
image domain	bin	bin	bin	int	int	int	bin	bin	bin	int
feature vector	proj	proj	proj	LBP	LBP	LBP	TriBP	TriBP	TriBP	Virage
vector length	510	256	130	256	256	256	75	75	75	?
visible fraction F_v	0.96	0.92	0.90	0.38	0.62	0.54	0.60	0.70	0.60	0.82
visible position P_v	0.97	0.97	0.91	0.88	0.78	0.88	0.78	0.83	0.79	0.91
retrieval quality Q_r	0.96	0.94	0.91	0.63	0.70	0.71	0.69	0.77	0.70	0.86

hardly degraded when using only 10 % of the most important elements; this means that the length of the 3x3 texel statistics feature vector need not be longer than 25-50. To compare our results with a commercially available method we fed 37 dpi versions of our images into an Illustra database with Virage Datablade for similarity matching. The results of Virage Datablade for the copy test set are given as well and show that Virage does well for these graylevel images; its performance is better than the two 3x3 texel based methods, though not as good as the projection method. The best performance is obtained at the highest resolution for the projection method, at the lowest resolution for the LBP method and at the intermediate resolution for the Trigram method. The projection method only misses about 1 in 16 copy test-pairs in the top $\lfloor^2\log n\rfloor$ ranks, the LBP method misses 1 in 3 and the Trigram 1 in 4. Although the Trigram retrieves a higher visible fraction, their average position is lower than that of the LBP method which is the reason why the quality measures differ less than the visible fractions. From the length of the feature vectors one can conclude that the Trigram method obtains the highest visible fraction per feature and is therefore the best value for money method. One can also see that gaining 2 % retrieval quality with the projection method means doubling the length of the feature vector.

5.2 Similar pairs with projection, Virage Datablade and texels

For a test-set of 12 rather similar to less similar test-pairs the same methods and measures were derived as for the copy case. The results for the methods given in Table 2 shows the same relative ordering as in the copy case: projection best, trigrams better than LBP. A difference with the copy case is the clear optimum performance of all methods at the intermediate resolution level of 37.5dpi. Although the statistics in the similar test-case are poor, one can still say that the very good performance of the projection method degrades very graceful; that the LBP method either entirely misses similars or finds them at the top; that the Trigram method even performs better for similars than for copies! It is clear that this has to be checked with a much bigger test-set of similars. Again Virage Datablade performs as well for copies as similars; the projection results are now closer to the results of Virage; but still better at 37dpi.

Table 2. Similar image retrieval performance of projection and texel methods in top $\lfloor {}^2\log n \rfloor$ ranks: $T = 12$ similar pairs embedded in $n = 5570$ files

Resolution in dpi	75	37	18	75	37	18	75	37	18	37
image domain	bin	bin	bin	int	int	int	bin	bin	bin	int
feature vector	proj	proj	proj	LBP	LBP	LBP	TriBP	TriBP	TriBP	Virage
vector length	510	256	130	256	256	256	75	75	75	?
visible fraction F_v	0.83	1.00	0.83	0.33	0.43	0.33	0.67	0.73	0.58	0.79
visible position P_v	0.71	0.85	0.83	1.00	0.95	1.00	0.70	0.90	0.83	0.92
retrieval quality Q_r	0.77	0.93	0.83	0.67	0.69	0.67	0.69	0.81	0.70	0.86

5.3 Copy retrieval performance of symmetry grouped Trigrams

That using invariant groups of Trigram patterns reduces the strength of the method is shown by the results in table 3. None of the RIM or RM symmetric feature vectors and none of the weighing schemes (equal, antilinear or band-pass) outperform the band-pass version of the un-symmetric Trigram feature vector. Using RIM symmetry for binarized gradient images is overdone, because taking the gradient already makes the images intensity reversal invariant (positives and negatives have the same gradient magnitude image). A RM symmetric band-passed feature vector however is also outperformed by the band-passed version of the un-symmetric Trigram vector.

Table 3. Copy retrieval performance of symmetry grouped Trigrams in top $\lfloor {}^2\log n \rfloor$ ranks: $T = 13$ copy pairs embedded in $n = 5570$ files

Resolution dpi	75	18	75	37	18	37	18	37	18	75
feature vector	RIM	RIM	RM	RM	RM	RM	RM	RM	RM	Tri
weights	equal	equal	equal	equal	equal	bndps	bndps	antiln	antiln	bndps
vector length	51	51	102	102	102	30	30	102	102	75
visible frac. F_v	0.54	0.39	0.61	0.69	0.54	0.61	0.54	0.69	0.54	0.84
visible pos. P_v	0.74	0.68	0.85	0.72	0.80	0.80	0.63	0.83	0.75	0.82
retriev qual. Q_r	0.64	0.54	0.73	0.71	0.67	0.71	0.59	0.76	0.65	0.83

6 Concluding remarks

Although we defined severe performance measures based on the visible top $\lfloor {}^2\log n \rfloor$ images in a database of size n (in our case 12 out of 5570) we found a number of content-based comparison methods that performed well enough to be of practical use. One of them, the projection vectors approach almost always has the right image at the first or second position; but also the texel statistic

vectors give meaningful results in a clear majority of test-cases. When a similarity appears in the visible top $\lfloor^2\log n\rfloor$ ranks, it is mostly within the first 3, which means that one can even consider to show less than half the top $\lfloor^2\log n\rfloor$ images, thereby saving precious communication time while at the same time saving waiting time. Due to the precalculated ranking results and the storage of the top ranks of each image in a postgres database, our internet demo program starts displaying results within a few seconds, which is fast enough to encourage use of this retrieval facility. Anyone with access to internet and Netscape can run our image retrieval demo at http://ind156b.wi.leidenuniv.nl:2000/.

7 Acknowledgements

A number of Erasmus and Master students implemented parts of the methods, internet demo and postgres database. We would like to thank Hans Guijt, Christof Dratner, Bernard Zwisschenbrugger and Silvia Poles for their implementations of various methods, Yuri Lausberg for the first version of the internet demo and Davy Wentzler for the second version of the internet demo and the postgres database. Special thanks go to Rinie Egas who generated the Virage Datablade results. This project is supported with a grant from Philips in the Netherlands and by The Netherlands Computer Science Research Foundation as far as M.S.Lew is concerned.

References

[Huijsmans96a] Huijsmans D.P., Lew M.S.; Efficient content-based image retrieval in digital picture collections using projections: (near)copy location; in Proc. 13th ICPR Vienna 1996 vol III pp 104-108.

[Huijsmans96b] Huijsmans D.P., Poles S., Lew M.S.; 2D pixel trigrams for content-based image retrieval; in (eds)Smeulders A.W.W, Jain R.; Image databases and Multi-Media search; Proc.1th Int workshop IDB-MMS A'dam 1996 pp 139-145.

[Lew96] Lew M.S., Huijsmans D.P., Denteneer D.; Content-based image retrieval: KLT, projections or templates; in (eds)Smeulders A.W.W, Jain R.; Image databases and Multi-Media search; Proc.1th Int workshop IDB-MMS A'dam 1996 pp 27-34.

[Ojala96] Ojala T., Pietikainen M and Harwood D.; A comparative study of texture measures with classification based on feature distributions; Pattern Recognition vol 29-1 Jan 1996 pp 51-60.

[Wang90] Wang L., He D.C.; Texture classification using texture spectrum; Pattern Recognition vol 23 (1990) pp 905-910.

[Wood96] Wood J.;Invariant pattern recognition: a review; Pattern Recognition vol 29-1 Jan 1996 pp 1-18.