Spatiogram-Based Shot Distances for Video Retrieval

Adrian Ulges[1], Christoph Lampert[1], and Daniel Keysers[1]

IUPR Research Group, DFKI,
Kaiserslautern, Germany
{ulges}@dfki.de
http://www.dfki.de/iupr

Abstract. We propose a video retrieval framework based on a novel combination of spatiograms and the Jensen-Shannon divergence, and validate its performance in two quantitative experiments on TRECVID BBC Rushes data. In the first experiment, color-based methods are tested by grouping redundant shots in an unsupervised clustering. Results of the second experiment show that motion-based spatiograms make a promising fast, compressed-domain descriptor for the detection of interview scenes.

Experiment 1: Clustering

Run-ID	concept	NN - error rate (%)	clustering - error (%)
CH-L1	baseline: color histograms	33.5	53.0
CS1D-JSD	**our framework**	**13.4**	**42.3**
CS-JSD	**our framework**	**15.4**	**46.5**
CS1D-PROB	different similarity measure [1]	36.2	90.6
CS-PROB	different similarity measure [1]	44.2	87.2
WH-L1	color histograms over local windows	16.1	50.3
BVW-Harris	”bag-of-visual-words”	24.8	91.9

Experiment 2: Interview Detection

Run-ID	concept	Precision (%)	Recall (%)
MH-L1	baseline: motion histograms	49.0	71.4
MS1D-JSD	**our framework**	**69.0**	**84.3**
MS-JSD	**our framework**	**49.4**	**78.1**
MS1D-PROB	different similarity measure [1]	44.1	67.6
MS-PROB	different similarity measure [1]	46.3	65.7
WH-L1	motion histograms over local windows	59.7	81.9
CS1D-JSD	color-based method	44.0	10.5

1 Introduction

Video streams in the TRECVID 2006 Rushes Task show highly unstructured, raw production material with lots of redundant takes. This redundancy is due to manifold reasons, including multiple takes of the same action, shots of the same object from different perspectives, interviews of the same person in front of different backgrounds, or even shots from the same semantic categories (like "still shots of gray photographs").

To support the video production process with automated retrieval tools, we cannot fall back on metadata, since no manual annotation has taken place, and usually neither speech nor screen text are present. Instead, visual content must be employed to identify redundant shots. A standard approach to such Content-based Video Retrieval is to approach the problem in four steps:

1. **Preprocessing**: segment the video stream into shots
2. **Feature Extraction**: extract preferably compact and discriminative features to reduce the dimensionality of the data
3. **Distance Measures**: measure the similarity of shots using a distance measure over the extracted features
4. **Matching**: use the distance measure to structure the shots, e.g. by clustering or classification

If embedded in this context, our work focuses on the feature extraction and distance measure steps. More precisely, we present a novel combination of features and distances. Our approach is based on *spatiograms*, a descriptor that extends histograms with a spatial component. Spatiograms were presented by Birchfield and Rangarajan [1] in the context of video tracking, but have not been used as a global descriptor for video before. As our distance measure between spatiograms, we suggest the Jensen-Shannon divergence, a distance measure that is well-motivated by information theory and broadly used.

To validate the performance of our framework, we have conducted two experiments on manually labeled video data: redundant shot clustering, and interview detection.

In the remainder, we first introduce our approach (Section 2) before we present experiments on shot clustering (Section 3) and interview detection (Section 4).

2 Our Approach

Our video retrieval framework is based on a novel combination of features and shot distance measures. In the following, we will first introduce the features – namely, spatiograms –, and after this discuss the distance measure and some of its properties – the Jensen-Shannon divergence.

2.1 Spatiograms

Birchfield and Rangarajan [1] have extended the popular concept of histograms with spatial layout information, yielding *spatiograms*. Like histograms, spatiograms are approximations of attribute distributions. Unlike histograms, the spatial layout for each attribute bin is part of the model as well. Promising results have been obtained with the tracking of local features in video, but spatiograms have not been used as a video descriptor on a global frame level before.

Given a video shot with an attribute $A(x)$ over positions X in the video stream (typical attributes are pixel color, texture, or motion), a histogram is defined by partitioning the attribute range into bins

$a_1, \ldots, a_n$ and counting the number of occurrences per bin:

$$h(a_i) = \frac{|\{x|A(x) \in a_i\}|}{|X|}.$$

If A is seen as a random variable, h can be used as a discrete approximation of the distribution of A: $P(a_i) \approx h(a_i)$.

One fundamental weakness of histograms is that the spatial structure of an attribute in an image or video is neglected. For example, the two scenes illustrated in Figure 1, though very different, have similar color histograms.

For spatiograms, the histogram model is replaced with a joint distribution of attribute value and location:

$$p(a_i, x) = \underbrace{P(a_i)}_{histogram} \cdot \underbrace{p(x|a_i)}_{spatialcomponent}$$

The open question left is how to model $p(x|a_i)$, the spatial distribution of an attribute bin. To fully preserve this spatial information require to store the complete image (or a quantized version of it, respectively). Other popular models like Gaussian mixtures would require less parameters, but therefore a clustering process over the support set of a bin $X_{a_i} := \{x|A(x) = a_i\}$. Instead, spatiograms use a simpler and faster descriptor given by a lower-order approximation in form of the intra-bin mean and covariance (i.e. a Gaussian single density):

$$\mu_{a_i} = \frac{1}{|X_{a_i}|} \sum_{x \in X_{a_i}} x, \qquad \Sigma_{a_i} = \frac{1}{|X_{a_i}|} \sum_{x \in X_{a_i}} (x - \mu_{a_i})^T (x - \mu_{a_i})$$

and we approximate

$$p(x|a_i) = \mathcal{N}(x; \mu_{a_i}, \Sigma_{a_i}) \tag{1}$$

Figure 1 illustrates the Gaussian spatial information for two color bins (blue and brown) in form of ellipses. Those capture the fact that colors appear in different parts of the frames, and allow to distinguish both scenes well.

Obviously, equation (1) is a strong simplification. A Gaussian distribution is a suitable model for the spatial layout of an attribute only if the pixels of a bin are concentrated in blobs and not spread over the whole image. Nevertheless, our experimental results show that this approximation is sufficient to significantly outperform the standard histogram model.

2.2 Distance Measure – Jensen-Shannon Divergence

Using spatiograms as features, the next issue is a "good" distance measure between them. For histograms, a wide variety of distance measures has been suggested [6], ranging from heuristic L^p norms to costly cross-bin measures like the Earth Mover's distance (EMD). The fundamental problem with these approaches is that it is not obvious how to extend them to spatiograms, i.e. how to integrate the spatial information.

Instead, our distance measure is based on the *Kullback-Leibler Divergence* (KL divergence), a well-motivated and widely used distance measure from information theory. Also, it has been shown to be optimal for retrieval in a probabilistic sense [10]. The most important argument, however, is that the KL divergence can be extended to spatiograms in a straightforward way.

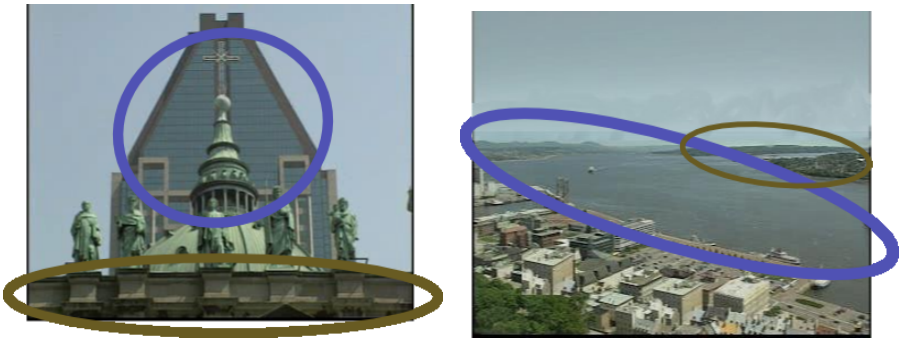


Fig. 1: Two different sample scenes. While their color histograms are very similar, the spatiograms capture the different color layout well and allow to distinguish both shots from each other.

To do so, we start from the definition of the KL divergence as a dissimilarity measure between two distributions P and P':

$$D_{KL}(P||P') = \int_a P(a) \cdot \log \frac{P(a)}{P'(a)} \, da,$$

where for discrete random variables the integral reduces to a sum (e.g., to a sum over all bins of a histogram).

To integrate spatial information into our similarity framework, we replace the histogram values $P(a_i), P'(a_i)$ with the spatiograms $p(a_i, x), p'(a_i, x)$, obtaining:

$$\begin{aligned} D_{KL}(p||p') &= \sum_i \int_x P(a_i)p(x|a_i) \cdot \log \frac{P(a_i)p(x|a_i)}{P'(a_i)p'(x|a_i)} dx \\ &= \sum_i P(a_i) \left(\log \frac{P(a_i)}{P'(a_i)} + D_{KL}(\mathcal{N}(.; \mu_{a_i}, \Sigma_{a_i})||\mathcal{N}(.; \mu'_{a_i}, \Sigma'_{a_i}) \right) \end{aligned}$$

with the KL divergence between two Gaussians [4]

$$D_{KL}(\mathcal{N}(.; \mu, \Sigma)||\mathcal{N}(.; \mu', \Sigma')) = tr(\Sigma \Sigma'^{-1}) - S + (\mu - \mu')^T \Sigma'^{-1}(\mu - \mu').$$

S denotes the dimension of the underlying space (in our case $S = 2$).

Note that the KL divergence is not symmetric. In our framework we follow common practice and use the Jensen-Shannon divergence (JSD) as a smoothed and symmetrized version instead.

$$D_{JSD}(p, p') = D_{KL}(p|\hat{p}) + D_{KL}(p'||\hat{p})$$

with $\hat{p} := \frac{1}{2}(p(a, x)+p'(a, x))$. To obtain a closed-form solution, we approximate the mixture $\frac{1}{2}(\mathcal{N}(x; \mu, \Sigma)+\mathcal{N}(x; \mu', \Sigma'))$ by a Gaussian $\mathcal{N}(x; \hat{\mu}, \hat{\Sigma})$ with

$$\hat{\mu} = \frac{1}{2}(\mu + \mu'), \quad \hat{\Sigma} = \frac{1}{2}(\Sigma_1 + \Sigma_2 + \mu_1^T \mu_1 + \mu_2^T \mu_2) - \hat{\mu}^T \hat{\mu}$$

Fig. 2: Several types of redundancy in our dataset: same person/scene/semantics

2.3 Matching

Given a set of video shots $s_1, .., s_n$, applying a distance measure to the corresponding video shots yields an $n \times n$ matrix $\mathcal{D}$ on which we can base a matching process. In a supervised context, such a matching can take place in form of Nearest Neighbor classification, or by inducing a kernel with $\mathcal{D}$ [4]. In an unsupervised setup, $\mathcal{D}$ can be used to group the shots using a (generally nonmetric) clustering.

3 Clustering Experiment

The first experiment to validate the performance of our framework is the estimation of *redundancy* in a set of video shots. Such redundancy may occur in various manners (Figure 2 illustrates some examples):

- shots of the same scene with different camera settings / from different perspectives
- shots showing the same person
- multiple takes of the same action
- shots with a common semantics (e.g., *shots in a library*)

We view the identification of redundant shots as a clustering problem: redundant shots should be judged as ”similar” by our distance measure and thus assigned to the same cluster.

3.1 Data

We have created a dataset of redundant shots grouped into clusters by manual labeling. Shot segmentation over a subset of the *BBC Rushes Test data* was done using the XViD[1] I-Frame detection. 149 shots, each of $4 - 40$ sec., were manually selected and labeled to obtain a set of 33 clusters showing various types of redundancy. Examples from the dataset are illustrated in Figure 2.

[1] www.xvid.org

3.2 Runs

Our runs in this experiment focus on features based on the *color* attribute. We test our approach as well as several baseline methods:

`CH-L1` - *Color Histograms* A global HSV Color Histogram is extracted for each shot. For means of efficiency, we do not take all frames into account, but update the histogram with all pixel positions in every 25th frame. 8 bins are used for the hue, 4 bins for the saturation, and 2 bins for the value (*H8S4V2*). All binning – in all of our runs – is regular. The histograms are matched using the $L1$ distance.

`CS-JSD` - *Our Framework* Like for the histograms, we use *H8S4V2* color spatiograms updated every 25 frames. The JS divergence as described in Section 3.2 is used as a distance measure.

`CS1D-JSD` - *Our Framework* To reduce the number of free parameters in our system (for each bin a_i, we store 6 values, i.e. its probability $P(a_i)$ as well as its spatial mean μ_{a_i} and its $2D$ symmetric covariance Σ_{a_i}), we found it useful to break down the 3-dimensional spatiograms into three one-dimensional ones. The number of bins thus reduces from $8 \cdot 4 \cdot 2 = 64$ to $8 + 4 + 2 = 14$.

The overall shot JSD is obtained by summing up the single channel JSD divergences, which can be motivated by the fact that the KL divergence for product distributions of independent variables reduces to the sum of the KL divergences for the single variables (thus, we approximately assume that H, S, and V are independent).

`CS-PROB / CS1D-PROB` - *Another Spatiogram Similarity Measure* Birchfield and Rangarajan [1] suggest an alternative similarity measure between spatiograms given by:

$$S_{PROB}(p, p') = \sum_i \sqrt{P(a_i)P'(a_i)} \cdot \mathcal{N}(\mu'_{a_i}; \mu_{a_i}, \Sigma_{a_i}) \cdot \mathcal{N}(\mu_{a_i}; \mu'_{a_i}, \Sigma'_{a_i})$$

We turn this similarity measure into a distance $D_{PROB}(p, p') = \exp(-S_{PROB}(p, p'))$.

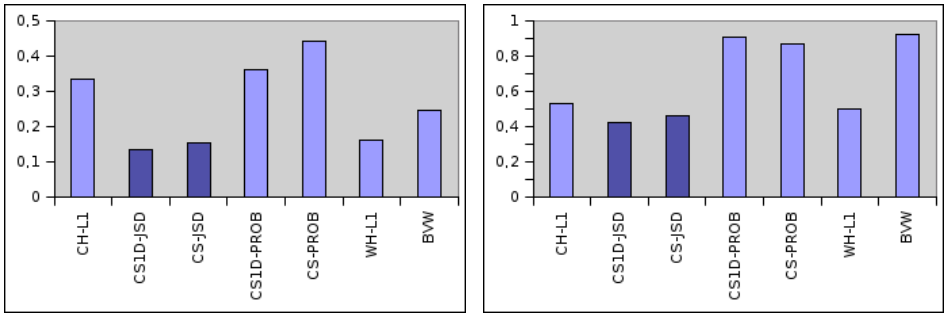
For the second run `CS1D-PROB`, the HSV spatiograms are decomposed into three 1D-histograms as described above for our framework, and the single distances are summed up.

`WH-L1` - *Local Histograms* This baseline method follows another popular idea to integrate spatial information (e.g., [8]), namely to subdivide the image into windows and then store a separate $H8S4V2$ color histogram for each window. Then, the $L1$ distance of the resulting feature vector is used. We use 2×2 windows, obtaining a number of parameters comparable to our spatiograms. In contrast to our framework, the approach is capable of representing non-Gaussian spatial layouts. On the other hand, it suffers from additional binning effects in the spatial domain.

`BVW-Harris` - *Bag of Visual Words* This approach follows an idea introduced by Sivic and Zisserman [7]: shot similarity is not based on global measures, but on the presence of the same categories of local features, referred to as "visual words". According to this metaphor, a video is a "visual document", and standard methods from text retrieval can be applied.

We use Harris corners from each 100th frame as visual words and describe them by 52 low-frequency DCT coefficients over their 16×16 pixel local YUV surrounding (28 intensity coefficients, and 15 for both chroma components). These are clustered using the K-Means algorithm over a training set of 500.000 whitened patches, obtaining a codebook of 500 visual words.

As a similarity measure, we extract patch occurrence histograms from video shots and use the angle between them as a similarity measure.



(a) (b)

Fig. 3: Results of redundancy detection (results of our framework highlighted): error rates for NN classification, and for unsupervised clustering

3.3 Results

The cluster-labeled shots allow us to measure the shot clustering performance of a distance measure in a straightforward way. We test all runs in two experiments:

NN classification In a first experiment, we view redundancy elimination as a classification problem. Each of the 149 shots is removed from the dataset, and the remaining 148 shots are used as training samples for NN classification. The cluster size in our dataset is at least two, such that there exists at least one ”correct” neighbor for each shot. The resulting error rates are presented in Figure 3(a).

Clustering In the second experiment, we use the distance matrices $\mathcal{D}$ for an unsupervised average linkage clustering. The number of clusters is determined using a compactness criterion by Davies and Bouldin [2]. For each clustering result, we consider the manual labels as ground truth and measure recognition error rates. To map the clusters in our results to the ground truth clusters, we solve a bipartite graph matching problem using the Hungarian algorithm (the problem and the procedure are described in [5]). Results are illustrated in Figure 3(b).

Obviously, both results are strongly correlated. If a distance measure performs well in NN classification, it can be expected to yield better clustering results as well. The best performance in both cases is achieved by our framework (`CS1D-JSD`), with a NN classification error rate of 14.4 %. The only comparable baseline results are achieved by the window histogram approach `WH-L1` with 16.1 %. Also, it can be seen that the reduction of dimensionality leads to a slight improvement compared to `CS-JSD` (15.4 %).

Unsupervised clustering is a far more difficult problem, which is indicated by significantly higher error rates. Again, the best error rate of 42.3 % is reached by our framework (`CS1D-JSD`). The results were validated by visual inspection as well, indicating that the performance is far from optimal, and many redundant shots were not clustered correctly (e.g., of the redundancies illustrated in Figure 2, only the second one was identified correctly).

(a) (b)

Fig. 4: Two interview scenes showing a typical motion pattern, and a spatiogram over motion vectors in $[-10, 10] \times [-10, 10]$ capturing this information.

4 Interview Experiment

The second experiment deals with the detection of a semantic feature, namely whether a shot shows an interview or not. For such an interview detection task, the color attribute is not a good choice. Instead, we use *motion* in form of MPEG-4 motion vectors. These provide fast, compressed domain features that can be extracted in sub-realtime. Furthermore, Figure 4 illustrates why motion spatiograms make promising features for interview detection: interviews often show a certain *motion pattern*, with the interviewee's body remaining static and the head moving occasionally with gestures. Such motion can be represented well by motion spatiograms as illustrated in Figure 4(b).

In this experiment, we apply the features and similarity measures of our framework over motion vectors. The performance is validated in a nearest neighbor classification framework.

4.1 Data

We segmented a subset of the *BBC Rushes Development data* into shots using the XViD I-Frame detection. 1404 of the resulting shots were labeled with *"showing an interview"* or *"not showing an interview"*. Shots were split up into 702 training and 702 testing shots (there was no overlap of *scenes* between testing and training).

4.2 Runs

All runs in this experiment (except for one color-based method) use MPEG motion vectors extracted by the XViD codec, which estimates them using a standard procedure based on prediction vectors and discrete gradient descent over the block SSD [9].

`MH-L1` - *Motion Histograms* A motion histogram describing a shot was computed over all motion vectors of all P-Frames in a shot (this holds for any motion-based method presented in the following). Motion vectors were clipped to the range $[-8, 8]/[-5, 5]$ and binned into 7×7 bins. The histograms are matched using the $L1$ distance.

`MS-JSD` - *Our Framework* For the spatiograms, we use a clipping to $[-10, 10] \times [-8, 8]$ with 5×5 bins. Again, all motion vectors were used. The JS divergence was used as a distance measure.

`MS1D-JSD` - *Our Framework* For reasons outlined in Section 3.2, we split the $2D$ spatiograms into two $1D$-spatiograms, using a clipping to $[-8, 8] \times [-5, 5]$ and $4 + 4$ bins. As before, the overall JSD is obtained by summing up the JSDs for the single dimensions (the x and y spatiograms).

`MS-PROB` / `MS1D-PROB` - *Another Spatiogram Similarity Measure* In these runs, the spatiograms described for `MS-JSD` and `MS1D-JSD` were combined with the spatiogram similarity measure presented in [1] in the same manner as described in Section 3.2.

`WH-L1` - *Local Histograms* Like in the first experiment, we use histograms over local windows as a baseline method. In this run, we use 3×3 windows, obtaining a descriptor dimension higher than for our spatiograms. Vectors were clipped to $[-8, 8] \times [-5, 5]$ and binned into 7×7 bins.

`CS1D-JSD` - *Color-based Method* To compare the interview detection performance of motion to color, the best color-based result from the clustering experiment was used (see Section 3.2).

4.3 Results

For each run, 3-NN classification was performed on the resulting distance matrix $\mathcal{D}$. Since there are significantly (about seven times) more negative samples than actual interview shots, we present precision and recall instead of error rates (see Figure 5(a)).

The best performance was achieved by our framework (`CS1D-JSD`) with a recall of 84 % and a precision of 69 % (corresponding to an error rate of 8.2 %), slightly better than the window histogram approach. The results look promising and indicate that it is possible to detect features based on their characteristic motion pattern in spatiograms. Also, all motion-based methods clearly outperform the color-based approach (`CS1D-JSD`).

Visual inspection indicated a lot of misclassifications where people are present in the image, but no interview takes place (for a false positive see Figure 5(b)). This indicates the potential to integrate the approach as a fast prefilter for audio or more intricate approaches like face detection.

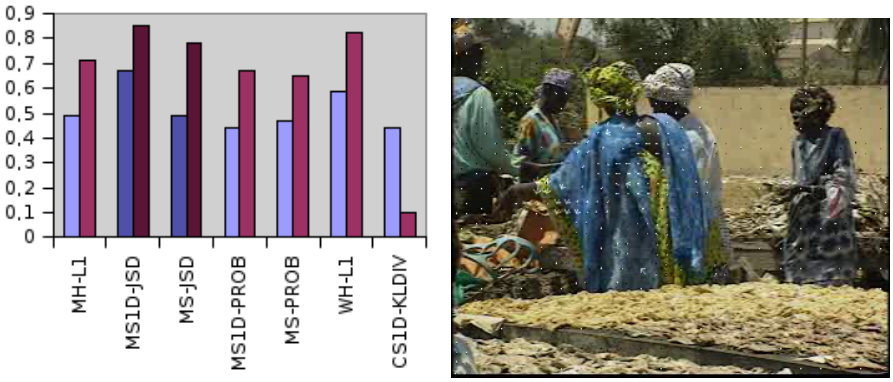
Also, it seems possible to learn further semantic features based on spatiogram motion patterns (e.g., if there are people present in the image, or if certain camera motions take place).

5 Discussion

In this paper, we have presented a shot matching framework based on spatiograms and the Jensen-Shannon divergence, which was found to outperform conventional baseline methods in two experiments.

In the redundancy detection experiment, however, none of the tested approaches has shown the capability to reliably structure shots of a complex dataset into meaningful clusters. Rather, the various types of redundancy demand specialized methods. For example, face recognition might be useful to detect shots of the same person. Localized approaches are particularly promising in this context and exist in much more elaborate form than tested here (`BVW-Harris`), baring scale, rotation, and affine invariance [3] and employing the spatial constellation of local features [7].

More promising are the results in the interview detection experiment, where our motion-based framework achieved good results (recall: 84%, precision: 69%) and indicates that spatiograms make an excellent fast-to-extract, compressed domain descriptor.



(a) (b)

Fig. 5: Results of interview detection: figures of precision and recall (results for our framework highlighted), and a typical misclassification

References

1. S.T. Birchfield and S. Rangarajan. Spatiograms versus Histograms for Region-Based Tracking. In *CVPR'05*, 2005.
2. D.L. Davies and D.W. Bouldin. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1(4):224–227, 1979.
3. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *CVPR'03*, 2003.
4. P. Moreno, P. Ho, and N. Vasconcelos. A Kullback-Leibler Divergence-Based Kernel for SVM Classfication in Multimedia Applications. In *NIPS'03*, 2003.
5. V. Roth, M. Braun, T. Lange, and J. Buhmann. A Resampling Approach to Cluster Validation. In *COMPSTAT'02*, 2002.
6. Y. Rubner. Perceptual Metrics for Image Database Navigation. Technical Report CS-TR-99-1621, Stanford University, 1999.
7. J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV'03*, 2003.
8. M.A. Stricker and A. Dimai. Color Indexing with Weak Spatial Constraints. In *SPIE'96*, 1996.
9. A.M. Tourapis. Enhanced Predictive Zonal Search for Single and Multiple Frame Motion Estimation. In *SPIE'02*, 2002.
10. N. Vasconcelos. On the Efficient Evaluation of Probabilistic Similarity Functions for Image Retrieval. *IEEE Trans. Inf. Theory*, 50(7):1482–1496, 2004.