

Content-based Image Retrieval via Subspace-projected Salient Features

Jyun-Hao Huang¹ Ali Zia¹ Jun Zhou^{1,2} Antonio Robles-Kelly^{1,2}

¹College of Engineering and Computer Science, ANU, Canberra ACT 0200, Australia

²National ICT Australia (NICTA)*, Locked Bag 8001, Canberra ACT 2601, Australia

Abstract

In this paper we present a novel image representation method which treats images as frequency histograms of salient features. The histograms are computed making use of linear discriminant analysis (LDA). The method employs saliency feature extraction and image binarisation. Then subspace-projected features are extracted. Using the saliency maps as the positive and negative labels, the image features are mapped onto a lower-dimensional space using LDA. This enables us to construct a 3D-histogram by direct binning on the feature space. This gives rise to a "cube" of image features which have been projected onto a lower-dimensional space so as to maximise the separability of the salient regions with respect to the background. Image retrieval can be performed by computing the distances between the histograms for the query image and the images in the database. We demonstrate our algorithm on a real-world database and compare our results to those yielded by codebook representation.

1. Introduction

Content-based image retrieval is an active research topic. Thanks to the success of digital imaging technology, retrieval from databases with a large number of images has attracted considerable interest from the computer vision and pattern recognition communities. However, despite the emergence of commercial systems such as QBIC (Query By Image Content) [10], FourEyes [14] and SQUID (Shape Queries Using Image Databases) [3], the retrieval of the best match in a dataset to a user-supplied query image based upon similarity remains an open problem.

A number of methods have been proposed so as to achieve robust and efficient systems for content-based image retrieval. These methods often represent an image as a

bag of features. A query image is then matched with images in the database by computing the distances between images. As a result, in general, object and image retrieval and classification techniques [23, 12, 17, 2] are based upon the summarisation of the image dataset using a codebook of visual words [4, 15, 11], which are used to retrieve images that best match the query. When a query image is provided by the user, the features in the image are compared with those on the codebook and a measure of similarity between the images in the dataset is computed so as to retrieve the closest match.

As a result, the design of an architecture for image retrieval requires both, an image representation suitable for search and a similarity measure that can be employed to rank the images with respect to the relevance to the query [19]. The main challenges in existing algorithms remain efficiency (in terms of speed and memory consumption), accuracy and simplicity. By efficiency we mean how quickly the result can be retrieved and how computationally costly is the image representation used. By accuracy we refer to the correctness of the images retrieved by the system provided a query image. Simplicity applies to the degree of ease of deployment and use of the algorithm.

Our main focus here is to present an image representation scheme which provides a good trade-off between accuracy and efficiency. In the proposed algorithm we try to achieve accuracy by using distinct image descriptors such as Harr features [20], edge and colour information. We have also used saliency to detect main components in the image. In achieving efficiency, we propose an image representation method using a 3D-space which can be viewed as a cube that gives rise to a feature histogram. This allows us to avoid the time consuming step of clustering to generate codebooks for images in the database.

The rest of the paper is organized as follows. Section 2 describes the details of the proposed method. In the section we elaborate further on the treatment given to the query and database images. The steps to construct the database include image binarisation using saliency maps, image feature extraction and dimensionality reduction by LDA. To process the query image, the same binarisation and feature

*NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

extraction steps are applied. The dimensionality of the extracted features for the query image is reduced using the subspace projection matrix recovered from the images in the database. We can then compare the query image to those in the database using nearest neighbor search. We report experimental results on the Oxford Flower Database in Section 3. Section 4 provides conclusions on the work presented here.

2. Content-based Image Retrieval

The fundamental idea of this research is to extract features from training images, train a model that best separates features from foreground and background using saliency maps and establish an image database using the model acquired. The same pre-processing and feature-extraction procedures are then applied to the query, whose similarity to the images in the database is measured using a nearest-neighbour approach. The diagram of the system is shown in Figure 1.

The first step in the algorithm is the saliency map extraction. Saliency applies to conspicuous areas in an image that appeal to human perception. By applying saliency detection [22, 5, 16], we can unclutter the image so as to separate the object of interest from the background.

Once the saliency map is at hand, we extract salient features from the images under study. These regions are recovered via binarisation, i.e. foreground-background segmentation, of the saliency maps. Three features are extracted. These are Harr features, edge and colour information. For the Harr features, we employ 12 filters in three orientations and four scales. For purposes of extracting edges, we have used a Sobel edge detector. The colour space used here is given by the CIE LAB gamut.

The features above are optimised using linear discriminant analysis (LDA). LDA and sub-space projection [21, 18] are commonly used techniques in image processing. We use Fisher LDA [8] to get the model which can efficiently discriminate the foreground and background features. This has two main advantages. Firstly, LDA reduces the dimensionality of the features, which in turn renders the image representation computationally efficient. Secondly, it achieves good separability between salient and inconspicuous features. This yields an image representation which is both, representative and compact.

Features from the LDA mapping are used to generate a 3D-histogram. Thus, we can view the feature histogram as a cube which represents each of the connected components in an image. These connected components correspond to salient regions and, therefore, each image is represented by a set of subspace-projected feature histograms. Image retrieval can then be performed by calculating the distances between histograms for both images. Here, we have used

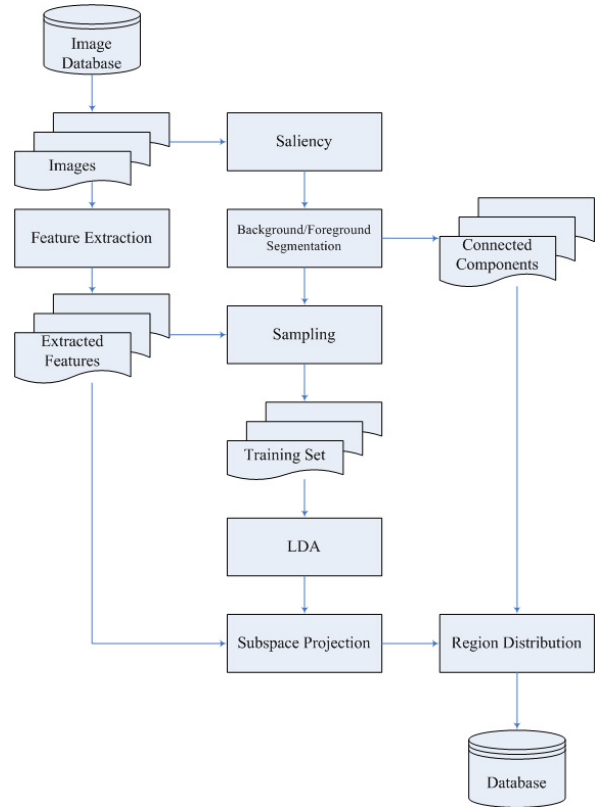


Figure 1. Flowchart of the training method. The database generation consists of five steps: feature extraction, saliency detection, foreground-background segmentation, subspace projection, and salient region representation.

the Euclidian distance [16] to measure how similar the histograms are to one another. Images that are represented by histograms that are closest to those in the query are retrieved as the best matchings.

This section is organised as follows. Section 2.1 elaborates on the feature extraction, saliency detection and foreground-background segmentation steps of the method. In Section 2.2, we turn our attention to the subspace projection step of the algorithm. Section 2.3 describes the process of matching a query image to those in the database.

2.1 Feature extraction and Saliency Detection

The objective of the feature extraction step is to map the query and database images into a feature space for purposes of recognition, classification, etc. As mentioned earlier, we have used three sets of image features, i.e. Harr-like, edge and CIE LAB. In Figure 2 we show the image features for an example image in the database. From left-to-right, we

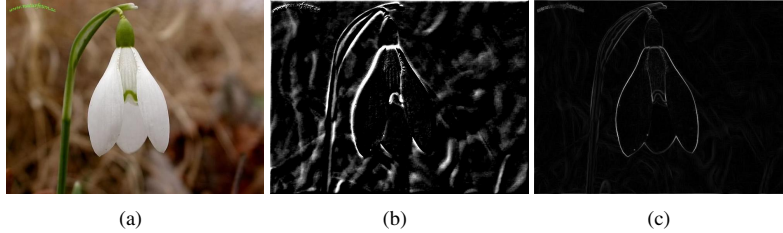


Figure 2. From left-to-right: Example image in the dataset, magnitude of the Harr-like wavelet filter response, and edge map yielded using the Sobel mask.

show the input image, the magnitude of the twelve Harr filter responses and the edges yielded by the Sobel mask.

For the construction of the image descriptors, we have used the Sobel detector response as a one-dimensional feature vector for the edge maps. Accordingly, the Harr descriptors are given by a 12-order vector containing the responses for each of the filters at every scale and orientation. In order to retrieve color features, images are transformed from RGB color-space to Lab color-space. As a result, the channels a and b are taken to form the color feature vector, which becomes a two-dimensional descriptor. It is worth noting in passing that the approach presented here is quite general in nature and permits the use of other local image descriptors, such as Gabor filters, local binary patterns, etc.

To develop our method, we note that, in general, an image often consists of a variety of features. Some of these features contain important information for image classification, whereas others are clutter that must be separated. This is intuitive from the input image, where the foreground flower is the main focus of the image. In the picture, the background, i.e. the blurred grassy foliage, is rather less important for recognition. This supports the observation that conspicuous areas in the image generally consist of regions with a high frequency or edges and a large amount of contrast information, whereas "flat" areas generally contain less meaningful information.

Based on this rationale, we aim at designing a filter which can separate high-frequency, i.e. contrast and edge information, from low-frequency components in the image, i.e. plain or blurred image regions. To this end, we build on the work in [6] and employ the "Cornersness" of pixels in the image as a measurement of their high-frequency information content. Recall that the "Cornersness", as presented in [6], for a pixel with coordinates u on the image lattice is defined as

$$R(u) = \det[M] - k(\text{trace}[M])^2 \quad (2.1)$$

where k is a constant and M is a 2×2 matrix computed from

image derivatives such that

$$M = \sum_{v \in \Omega_u} W(u, v) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \quad (2.2)$$

where I_x and I_y are the image derivatives at u in the x and y directions on the image lattice, i.e. row and column-wise, Ω_u is the set of pixels corresponding to a neighbourhood centered at u and W is a Gaussian kernel function

$$W = \exp\left(\frac{-(u^2 + v^2)}{2\sigma^2}\right) \quad (2.3)$$

with a variance σ .

As a result, we can view the matrix M as a weighted linear combination of the variation in the image, as captured by the partial derivatives, across a neighbourhood centered at the pixel of interest. In the left-most panel of Figure 3, we show an example of the $R(u)$ values for the input image in Figure 2. As can be seen from the figure, those pixels in the edge or high contrast regions of the image have higher "Cornersness" values. It is worth stressing that since, both, image derivatives and the matrix M can be computed efficiently using convolutions, this opens-up the possibility of using Equation 2.1 as a computationally inexpensive saliency detector.

Thus, in our method, we employ the values yielded by Equation 2.1 as a saliency map that can then be binarised so as to separate the foreground features from the background clutter. To provide comparison with respect to other saliency methods elsewhere in the literature, we have also implemented the saliency detector in [7]. This method extract saliency feature based on color, intensity and orientation information. As a result, points at which these three features change dramatically are considered to be conspicuous on the saliency map. As an example, in the third panel of Figure 3, from left-to-right, we show the saliency map yielded by the method in [7] for the input image in Figure 2. We can clearly see the difference between the two methods. In contrast with the method of Itti et al. [7], the method proposed here emphasizes the regions that contain edges and corners.

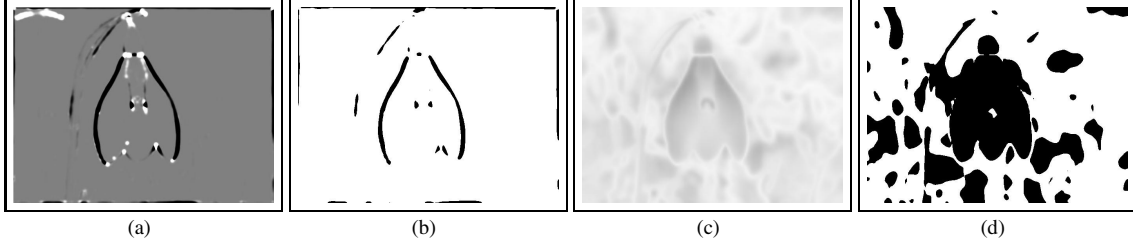


Figure 3. From left-to-right: First panel: Saliency map yielded by Equation 2.1; Second panel: Foreground-background segmentation for the saliency map in the left-hand panel; Third panel: Saliency map yielded by the method of Itti *et al.*; Fourth panel: Foreground-background segmentation on the Itti *et al.* saliency map.

2.2 Subspace Projection

For purposes of recovering the salient features for our content-based image retrieval application, we apply the method in [13] so as to find the optimal saliency threshold which separates the foreground from the background. This is, effectively, a binarisation task that permits the extraction of those connected components in the image which present high saliency values. The background/foreground segmentation for both, our approach and the alternative in [7] corresponding to the sample image in Figure 2 is shown in Figures 3(b) and (d).

With the extracted features and the binarisation result obtained from saliency detection, we can turn our attention to the subspace projection task. For the sake of efficiency, we have randomly sampled features from both background and foreground for the images in the dataset. We have done this since, as the number of training images increases, also does the number of features available for the recovery of the subspace projection matrix. Selected features are then labeled as either positive (foreground) or negative(background) ones and used as input to a linear discriminant analysis [1] as a means to dimensionality reduction. Recall that LDA aims at recovering the best projection subspace ψ^* such that

$$\psi^* = \arg \max_{\psi} \frac{\psi^T S_b \psi}{\psi^T S_w \psi} \quad (2.4)$$

where S_b and S_w denote between and within class covariance, respectively. For our two-class problem, these are defined as follows

$$S_b = (m_1 - m_2)(m_1 - m_2)^T \quad (2.5)$$

and

$$S_w = \sum_{i \in \{1,2\}} \sum_{d \in D_i} (d - m_i)(d - m_i)^T \quad (2.6)$$

where i denotes the class index, i.e. 1 for foreground and 2 for background, D_i is the data set for the class indexed i and m_i is the corresponding mean value.

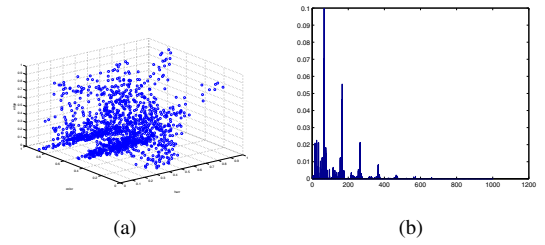


Figure 4. Left-hand panel: 3D mapping of the subspace projected image features; Right-hand panel: Frequency histogram for the 3D space in the left-hand panel.

Thus, LDA recovers a subspace projection matrix which minimises within class variance and maximises variation between classes. Note that, here, we have used three types of image features. Therefore, for each of these features, a projection matrix is recovered. That is, to each image descriptor it corresponds a subspace projection matrix.

Once the subspace projection matrices are at hand, we proceed to dimensionally reduce the features in each connected component so as to recover a frequency histogram that can be used for purposes of image retrieval. Recall that, in previous sections, we obtained a binary map which separates foreground from background in the scene. From this binary map, connected components in the foreground can be obtained. These connected components are, effectively, a mask that can be used together with the subspace-projected image features in order to construct a set of frequency histograms for each image. To this end, we map the features corresponding to each connected component in the image onto a 3D space as that in Figure 4 (a) and generate a frequency histogram, where each bin is a partition in the 3D space. As a result, for an image with K connected components, we recover K histograms. An example of such histograms is shown in Figure 4 (b).

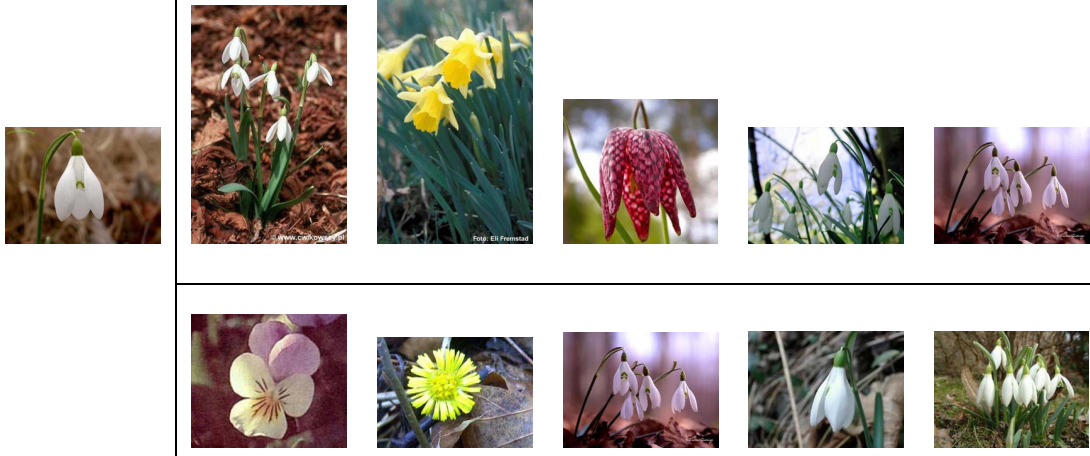


Figure 5. Example query results. Left-hand panel: Input query image; Top row, second to sixth columns: Results, ordered by relevance, yielded by the proposed method; Bottom row, second to sixth columns: Results, ordered by relevance, obtained using the saliency maps recovered by the method of Itti *et al.*. Both sets of results were recovered using frequency histograms computed from subspace-projected image features.

2.3 Query Process

When a query image is provided, we commence by extracting features as described in section 2.1. Then, subspace-projection is applied to the features extracted. Here, we use the subspace projection matrices obtained from the database as described in section 2.2. The subspace-projected image features are used to compute frequency histograms corresponding to each of the connected components in the query image.

After the query image has been processed and its image features transformed, we can then turn our attention to the comparison procedure. The histograms obtained are compared with those corresponding to the images in the database. This is done by computing the Euclidean distance between each histogram-pair. The distance τ is given by

$$\tau_{a,b} = \sum_{n \in N} (H_{a_n} - H_{b_n})^2 \quad (2.7)$$

where N denotes the number of bins contained in each histogram and H_{a_n} and H_{b_n} denote two bins corresponding to a pair of histograms.

Note that the query and the data images may contain different number of connected components. Hence, for purposes of retrieval, we employ the similarity measure τ^* defined as

$$\tau^* = \min_{i \in K_a, j \in K_b} \tau_{i,j} \quad (2.8)$$

where K_a and K_b are the numbers of connected component contained in the two images indexed a and b under comparison. Recall that, as mentioned before, here we use a nearest

neighbour retrieval scheme. Therefore, as per our results, the image in the database which has the smallest τ^* will be retrieved as the best match.

3. Experiments

In this section, we illustrate the utility of our method for content-based image retrieval. To this end, we make use of the Oxford flowers dataset [11]. This dataset has a training and a testing set of 680 and 340 images, respectively. Both, the training and testing sets contain 17 different species of flowers. Each specie has common features such as colour, shape and texture. Also, in all our experiments, we have set the number of bins for the frequency histograms to 12^3 , i.e. 12 bins per subspace-projected feature.

Firstly, we provide a qualitative evaluation of the results delivered by our method. To this end, we present two sets of example query results in Figure 5. In the left-hand panel, we show the input query image. In the top and bottom rows of the remaining columns we show the query results ordered by relevance from left-to-right. In the top row, we show the results yielded by our method. In the bottom row, we show the results delivered by our frequency histogram representation of the subspace projected features extracted using the saliency method in [7]. Note that, in the case of the results recovered using our method, the most relevant query result is, indeed, an image corresponding to the query specie. The results recovered using the alternative saliency map also deliver a number of alternative images corresponding to the query specie. Nonetheless, our algorithm delivers a margin of improvement in the order of relevance in the query

	1-NN	2-NN	3-NN	4-NN	5-NN
Saliency + Freq. Hist.	28.8%	41.5%	50.3%	58.5%	63.5%
Itti <i>et al.</i> + Freq. Hist.	32.4%	43.2%	49.4%	56.8%	61.2%
Itti <i>et al.</i> + Codebook	35.3%	44.1%	51.2%	58.2%	62.3%

Table 1. Performance comparison.

results.

Now, we turn our attention to a more qualitative analysis of the results. In table 1, we show the retrieval performance using a number of options for saliency detection and image representation. A number of alternatives in the literature [4, 15, 11, 23] make use of a codebook of visual words so as to represent the images under study. Similarly to our frequency histogram representation, the codebook approach permits the use of a nearest neighbour classifier for purposes of retrieval. As a result, we have used the codebook approach as an alternative to our histogram representation. Thus, in the table, we show results for the k -nearest neighbours, with $k = \{1, 2, 3, 4, 5\}$, recovered making use of the codebook obtained using the approach in [23]. Our histogram representation is computed using the saliency maps of Itti *et al.* [7] and the method presented throughout the paper, i.e. the frequency histograms for the subspace-projected features corresponding to the "Corner-ness" saliency maps.

From the table, we can conclude that the proposed saliency detector outperforms the alternative for $k \geq 3$. It is worth noting that, nonetheless for $k = 1, 2$ the results are comparable, our saliency map recovery is more computationally efficient. Our method also provides a margin of advantage for $k \geq 4$, with comparable results for other values of k . This suggests that the histogram representation presented here is an effective way of representing images in the dataset. Moreover, since the codebook approach requires clustering on a large dataset of image features, it can be computationally demanding. By making use of direct binning to generate the histogram, our approach is more computationally efficient. Another advantage of the proposed method is that LDA can greatly reduce the dimensionality of features. This, when compared to raw descriptors, such as SIFT [9], opens-up the possibility of greatly reducing database storage requirements and query run-times.

4. Conclusions

In this paper, we have presented a method for purposes of contents-based image retrieval which combines saliency detection, subspace projection and a compact image representation based upon frequency histograms. The saliency map extraction scheme presented here is computationally efficient and can be achieved via convolution routines. More-

over, the binarised saliency maps permit the use of foreground and background image features for purposes of recovering a subspace projection matrix making use of linear discriminant analysis. These subspace projection matrices allow us to construct frequency histograms in which the separability between background and foreground features is maximum. We have provided results on a real-world database and compared our method to an alternative representation often used in the literature.

References

- [1] I. Borg and P. Groenen. *Modern Multidimensional Scaling, Theory and Applications*. Springer Series in Statistics. Springer, 1997.
- [2] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Int. Conference on Computer Vision*, 2007.
- [3] S. A. M. Farzin and J. Kittler. Robust and efficient shape indexing through curvature scale space. In *Proceedings of the 7th British Machine Vision Conference*, volume 1, pages 53–62, 1996.
- [4] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Comp. Vision and Pattern Recognition*, pages II:524–531, 2005.
- [5] C. Harris and M. Stephens. On measuring low-level saliency in photographic images. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 84–89, 2000.
- [6] C. J. Harris and M. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference*, pages 147–151, 1988.
- [7] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.
- [8] X.-B. Li, J.-Y. Li, and R.-H. Wang. Dimensionality reduction using mce-optimized lda transformation. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pages 37–40, 2004.
- [9] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [10] W. Niblack et al. The QBIC project: Querying images by content using color, texture and shape. In *Proc. SPIE Conference on Storage and Retrieval of Image and Video Databases*, pages 173–187, 1993.

- [11] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1447–1454, 2006.
- [12] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Comp. Vision and Pattern Recognition*, pages II:2161–2168, 2006.
- [13] N. Otsu. A thresholding selection method from gray-level histograms. *IEEE Trans. on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.
- [14] R. W. Picard. Light-years from Lena: Video and image libraries and the future. In *International Conference on Image Processing*, volume 1, pages 310–313, 1995.
- [15] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modelling scenes with local descriptors and latent aspects. In *Int. Conference on Computer Vision*, pages I:883–890, 2005.
- [16] P. L. Rosin and G. A. W. West. Saliency distance transforms. *CVGIP: Graphical Model and Image Processing*, 57(6):483–521, 1995.
- [17] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Int. Conference on Computer Vision*, pages II:1470–1477, 2003.
- [18] F. Tang and H. Tao. Fast linear discriminant analysis using binary bases. In *Proceedings of the 18th International Conference on Pattern Recognition*, volume 2, pages 52–55, 2006.
- [19] N. Vasconcelos. On the efficient evaluation of probabilistic similarity functions for image retrieval. *IEEE Trans. on Informaiton Theory*, 50(7):1482–1496, 2004.
- [20] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of CVPR*, pages 511–518, 2001.
- [21] J. Ye. Least squares linear discriminant analysis. In *Proceedings of the International Conference on Machine Learning*, pages 1087–1094, 2007.
- [22] R. B. Yossi Cohen. Inferring region saliency from binary and gray-level images. *Pattern Recognition*, 36(10):2349–2362, 2003.
- [23] J. Zhou and A. Robles-Kelly. A quasi-random sampling approach to image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.