APPEARANCE INDEXING

Guoping Qiu

School of Computer Science, The University of Nottingham, United Kingdom

ABSTRACT

Although it is very hard to quantify the visual impression of an image, we conjecture that the overall visual appearance of an image may be the combined effect of the appearances of image patches of various shapes and sizes. We can imagine that all possible appearances of image patches of all possible shapes and sizes form a conceptual appearance space. Each point in the appearance space therefore corresponds to a certain combination of the object parameters (shape, size, surface texture, pose, orientation, etc) and the imaging conditions (illuminating source, viewing angle, sensor response etc.). Using real sensor data and unsupervised learning, statistically most representative appearance prototypes can be found to approximate the appearance space. Statistics of these appearance prototypes present in an image therefore characterize the image content, which in turn can be used to perform tasks such as content-based image retrieval.

1 INTRODUCTION

At the lowest level, computational vision algorithms operate on intensity images. In recent years, a class of vision technique that derives representation, recognition and interpretation models directly from image intensities has received much interest. These techniques are generally referred to as appearance-based vision [10-19]. In contrast to techniques based on image segmentation, figure ground separation, and shapes modeling (all very hard), appearance-based techniques build models directly from image data and thus may provide a quicker, simpler route to the development of vision systems.

In [10] statistics of single pixel appearance (color) was used for object recognition and localization. In [11, 12, 17], appearance of image patches was used for object recognition. In [13, 14], the appearance of the entire image was used for object recognition. Appearance models can not only be applied to recognition, they have also been applied to obstacle detection for mobile robots [18] and 3D-object synthesis [16] for computer graphics.

Appearance-based techniques have also been applied to content-based image indexing and retrieval. The best known example is the color indexing method [10]. Several researchers including the author have tried to characterize small image patches using vector quantization (VQ) for image indexing and retrieval [22 - 27]. While others try to directly exploit classic VQ based image coding [29], our effort has been directed toward the development of efficient representations for color image indexing and retrieval [22 –24].

In this paper, we extend our earlier work to multiresolution patches. Because such method is intrinsically linked to the idea of appearance-based vision, we refer to such method appearance indexing. While color indexing treats individual pixels independently, appearance indexing operates on a group of neighboring pixels and attempts to characterize the combined appearance of multiple pixels simultaneously. We make use of a conceptual “appearance space”, which can be regarded as analogy to color space in color indexing. A set of appearance prototypes, learned from real image data, is used as statistically most representative appearances in the world. Appearance indexing amounts to compile statistics (probabilities) of these appearance prototypes present in the image. To illustrate the usefulness of appearance indexing, we have applied it to content-based image retrieval and we will present experimental results to demonstrate its effectiveness.

2 APPEARANCE QUANTIZATION AND INDEXING

Vision deals with brightness images that are functions of a variety of variables, including surface reflectance, illumination, imaging geometry and sensor parameters. The appearance of an object therefore depends on these many variables. Clearly, it would be impossible to build any model that can capture all the varying factors that affect the appearance of the object. Furthermore, these many variables intertwine to produce the final pixel values, which are the only readily available data. Therefore, it is extremely difficult, if not impossible, to derive analytical models explicitly. Appearance based vision uses pixel values directly to build recognition and interpretation models, therefore the most often used (probably most suitable) methods are learning techniques [14].

An effective, efficient, and suitable representation is the key starting point to building computer vision systems [28]. Appearance representation scheme plays a critically important role in appearance based vision. From a practical implementation point of view, the representations should be in some low dimensional space to make the computational task feasible. Eigen analysis or principal component analysis used in [13, 14, 17] captures statistically most significant varying factors to enable computation to be performed in the much lower dimensional eigenspace. The purposes of using filtering in [11, 12] are similar, i.e., to represent (capture) the appearance in a lower dimensional space to facilitate computation. Both the eigenspace and the filtering outputs are continuous. In order to build recognition models (represented in computer bits), they have to be discretised.

The aggregated effect of illumination, surface reflectance, object shape, orientation, pose, and view point, etc., are reflected in the pixel intensities, which in turn form the appearances. If we view all possible appearances as discrete points in the “appearance space”, then each point in that appearance space captures (represents) a certain combination of the object shape, orientation, pose, surface reflectance, illumination, imaging geometry, and sensor characteristics, etc. Obviously, in general, there is infinite number of such combinations. Fortunately, in real world application environments, such combinations are finite. Of course, from a computational point of view, it would be impractical to store all possible combinations, even though they are finite. However, in many cases, certain representation imprecision can be tolerated and the whole feature space (in our case appearance space) can be approximated by a small number of prototype features (appearances).

The only knowledge we have is pixel intensities. We do not know what causes the pixels appear the way they do. The same color can be a result of many different combinations of illuminants and surfaces, or two different colors can come from the same surface under different lighting. Therefore, in order to find such appearance prototypes, an unsupervised learning technique has to be used. In unsupervised learning, the algorithms must discover for themselves patterns, features, regularities, correlation, or categories in the input data [9]. In the context of VQ, the appearance prototypes are the codebooks of the VQ coder, and thus can be designed using a number of well-studied codebook design methods [29].

Researchers have been trying to describe the appearance at pixel, regional and global levels. Global features cannot deal with occlusion and pixel level features do not contain spatial information. Therefore, we would like to model regional appearances (with the global and pixel level appearances as special cases). To increase discrimination power, we would also like to model the appearance at multiple resolution [30]. Basically, there are two ways to form an appearance vector for a region. One is to use all the raw pixel values to form a vector, and this is the Eigen analysis approach [13, 14]. The other is to drive another set of number using some form of local operators [11, 12]. We would also like to include color. Straightforward extension to color is to treat each channel as a gray scale image. A more efficient approach would be to use some form of opponent color space [2] and take into account the bandwidth and other properties of different channels [24-26].

By using local operators such as derivatives of Gaussian and Gabor, the appearance vectors can be made robust to scale changes [11, 12]. However, such vector cannot represent colors of the objects explicitly. In the cases where color is a useful cue, additional vectors, perhaps based on raw pixel intensities (colors) will have to be used. Using the raw pixel intensity to form the appearance vector has the advantage of including color explicitly but such vector may be sensitive to large scale variations. A combination of the two approaches will probably be desirable in some applications. Because these appearance vectors will be sent to a learning module, robustness can also be achieved through learning if large enough example data sets are available.

Once we have chosen an appropriate appearance vector representation scheme, we can use a number of established methods to design the prototypes or codebooks (we shall use codebook and prototype interchangeably). A vector quantizer is described by an encoder Q, which maps the k-dimensional input vector X to an index $i \in N$ specifying which one of a small collection of reproduction vectors (codewords) in a codebook C = {C$_i$; i ∈ N} is used for reconstruction. Although we do not use it in this work, there is also a decoder, Q^{-1}, which maps the indices into the reproduction vectors, i.e., X' = $Q^{-1}(Q(X))$. A winner-take-all competitive learning algorithm [9] or other clustering methods [29] can be used to find the prototypes. We have found the frequency sensitive competitive learning rule [8] worked robustly and gave excellent performance.

The appearance prototypes learned from training samples should capture statistically most representative appearances in the appearance space. Each prototype should reflect the characteristics of certain aspects of the appearance parameters. Statistics of these appearance prototypes in a given image should therefore reflect the content of the image. Similar to color indexing, a histogram of appearance can be constructed to characterize an image. From a given image, many appearance vectors ξ can be created, and appearance histogram $\mathbf{H} = \{h_i \mid i = 1, 2, \ldots N\}$ is defined as in (1) and we call it appearance indexing. Therefore, the i^{th} element of the appearance histogram h_i is the probability that an appearance vector from the image is most closely approximated by the i^{th} appearance prototype.

$$h_i = \Pr(Q(\xi) = i), \forall i \qquad \textbf{(1)}$$

3 An Implementation

We present one possible implementation of appearance indexing based on our earlier work [22-24]. In this scheme, we first decompose a given image into multilevel Gaussian pyramid [7]. At each level, the image is represented in an opponent color space. The appearance vectors are obtained directly from raw pixel values (we will present results of using local operators in another application). The appearance prototypes are created by training an unsupervised neural network. Let $\{\boldsymbol{I}_l\,(\mathrm{x}, \mathrm{y})\} = \{r_l\,(\mathrm{x}, \mathrm{y}), g_l\,(\mathrm{x}, \mathrm{y}), b_l\,(\mathrm{x}, \mathrm{y})\}$ be the l^{th} level image in an image pyramid. For implementation simplicity, we use the Burt Adelson Gaussian pyramid [7].

These images are then transformed into an opponent space [2]. We use the YCbCr space [4] (similar spaces [3] can be used and we have observed similar results). At each level, image patches (blocks) of m x n pixels, are formed. Let $\{\boldsymbol{B}_l\,(i, j)\} = \{Y_l\,(i, j), Cb_l\,(i, j), Cr_l\,(i, j),\} \mid i = 1, \ldots m, j = 1 \ldots, n\}$ be an image patch at level l. For each block, we form two appearance vectors as follows

$$A_l = \left\{ \frac{Y_l(i,j)}{M_{B_l}} \middle| , \forall i, j \right\}$$

$$C_l = \left\{ \frac{Cb_l(2i,2j)}{M_{B_l}}, \frac{Cb_l(2i,2j)}{M_{B_l}} \middle| \; i = 1,2,\ldots\frac{m}{2}, j = 1,2,\ldots\frac{n}{2} \right\} \qquad \textbf{(2)}$$

$$where \quad M_{B_l} = \frac{1}{m \times n} \,{}^{m}_{i=1}\,{}^{n}_{j=1}\, Y_l(i,j)$$

$\boldsymbol{A}_l$ is the achromatic appearance vector, and $\boldsymbol{C}_l$ is the chromatic appearance vector of block $\boldsymbol{B}_l$. From a signal analysis's point of view, ***Cb*** and ***Cr*** have lower bandwidth than ***Y***, and from a human vision perspective, human vision system is less sensitive to distortions in the opponent channels [2]. Therefore, $\boldsymbol{C}_l$ is formed by concatenating sub-sampled ***Cb*** and ***Cr*** pixels. The block average of the ***Y*** component, which captures most of the energy of the block, is used to normalize the pixels. This has the effect of making the appearance vectors less sensitive to absolute pixel intensities. Instead of forming one appearance vector using all pixels from all channels, we form two separate appearance vectors. From a computational point of view, we can work on two lower dimensional vectors instead of one very high dimensional vector, thus avoiding the curse of dimensionality problem. From a human vision's point of view, color and pattern are separable, and there probably exist separate channels in the human visual system to process spatial pattern and color [1]. Clearly, spatial patterns (texture and other spatial changes) are mostly captured in $\boldsymbol{A}_l$, and $\boldsymbol{C}_l$ captures the chromaticity of the block.

Because the image has been decomposed into multilevel Gaussian pyramid and we extract patches from every level, we are therefore in effect extracting multi-resolution appearance vectors. An m x n patches at level l corresponds to a $2^l m$ x $2^l n$ patch in the original image. For an L-level (including the original) pyramid, the appearance vectors are obtained from block sizes of m x n, $2m$ x $2n$, $4m$ x $4n$, …, $2^{L-1}m$ x $2^{L-1}n$.

Once we have the appearance vectors, appearance prototypes (VQ codebook) can be created using a simple unsupervised neural network [8] or other clustering techniques [29]. Although different sizes of patches can be used, we use 4 x 4 pixel patches for its moderate computational complexity, and which will cover areas of 4 x 4, 8 x 8, 16 x 16, 32 x 32, … in the original image. We have used over 15 million patches of 4 x 4 pixels obtained from natural color images to create a 256 achromatic and 256 chromatic appearance prototypes. These prototypes can form 64K (256 x 256) 4 x 4 patches. It is also important to note that these appearance patterns will be used at various levels of the pyramid, therefore, we have appearance prototype patterns of multiple resolutions.

With these prototypes, appearance indexing can be formed. We can form an achromatic appearance histogram $\boldsymbol{H_A}$ by indexing the achromatic appearance prototypes. We can also form a chromatic appearance histogram $\boldsymbol{H_C}$ by indexing the chromatic appearance prototypes. A joint histogram $\boldsymbol{H_{AC}}$ of indexing the achromatic and chromatic appearance prototypes can also be formed.

4 Application to Image Retrieval

Content-based image retrieval is currently an actively research area in which computer vision techniques can play a useful role [21]. Appearance indexing is suitable for such applications. It is an extension to the classic color indexing (setting the block size to 1 pixel and operating only on 1 level pyramid, appearance indexing becomes color indexing). In this section, we present experimental results of applying appearance indexing to content-based image retrieval.

The database used in our experiment consisted of 20,000 color images from the commercially available Corel color photo collection. To build the database, each image was decomposed into a 3-level Gaussian pyramid. At each level, an achromatic appearance histogram and a chromatic histogram were formed by taking appearance vectors from 4 x 4-pixel non-overlapping blocks. Let $\boldsymbol{HP}_{Al}$ $\boldsymbol{HP}_{Cl}$, $\boldsymbol{HQ}_{Al}$, and $\boldsymbol{HQ}_{Cl}$ be the achromatic and chromatic histograms of images $\boldsymbol{P}$ and $\boldsymbol{Q}$ respectively, the similarity of two images is measured as

$$D(P,Q) = \underset{\forall l}{sum}\left|HP_{Al} - HQ_{Al}\right| + \underset{\forall l}{sum}\left|HP_{Cl} - HQ_{Cl}\right| \qquad \textbf{(3)}$$

We constructed a query database consisting of 69 classes and a total of over 400 query images. Each class consisted of various numbers of similar images, and two examples are shown in Fig. 1. These images were then embedded into the 20,000-image database. Using each of these images as query, those in the same class are used as the correct answer.

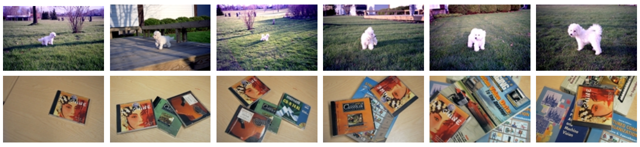


Fig.1, Example of two classes of query images

Let $\boldsymbol{Q}_i$ be the ith query image and $\boldsymbol{Q_i}(1)$, $\boldsymbol{Q_i}$ (2) … $\boldsymbol{Q_i}$ (N_i) be the N_i "correct" answers to the query $\boldsymbol{Q}_i$. We define the accumulated recall (AR) and accumulated precision (AP) as (4). Fig. 2 shows the AR(k) and AP(k) performances of the single level appearance indexing (AI) [24] and the new 3-level AI methods. Fig. 3 shows the AR(k) and AP(k) performances of a 3 level AI method and a version of the color correlogram (CC) method (64-color, 4-distance as in [5]). Fig. 4 shows an example of returned images for a query. It is seen that multiresolution AI give better performance than single resolution AI and that the AI technique gives a better performance than the CC method.

$$AR(k) = \underset{\forall i}{sum}\left(\left|\{Q_i(j) \mid rank(Q_i(j)) < k\}\right| \big/ N_i \right) \qquad \textbf{(4)}$$

$$AP(k) = AR(k) \big/ k$$

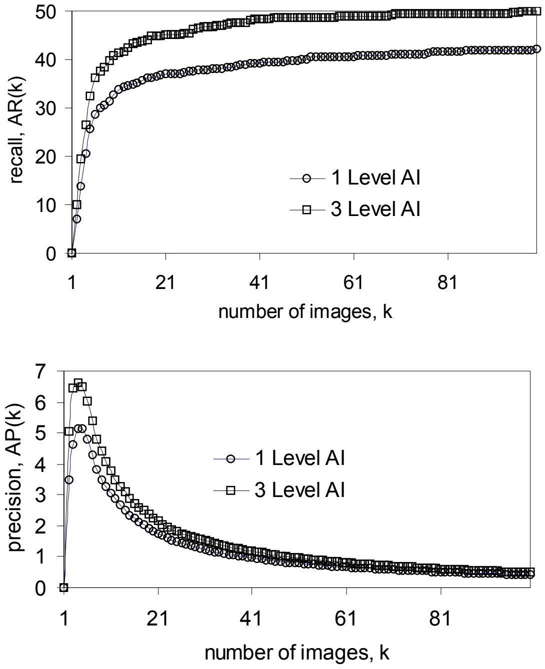


Fig. 2, Recall and precision performances of single- and multi- resolution AI accumulated over 60 queries.

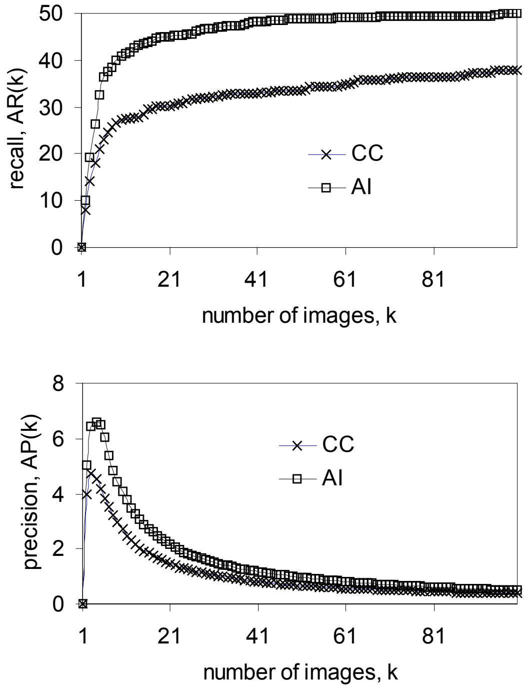
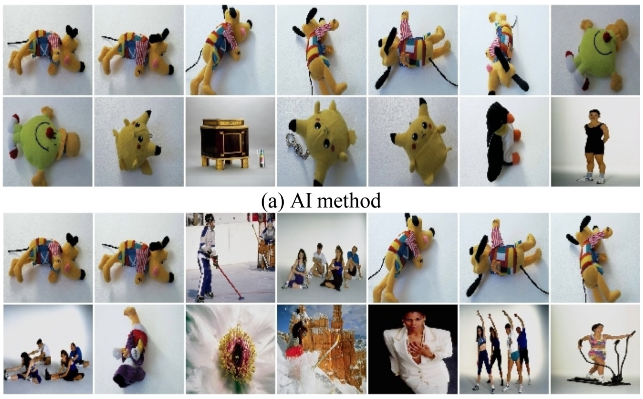


Fig. 3, Recall and precision performance of the new AI method and the CC method accumulated over 60 queries.



(a) AI method

(b) CC method

Fig. 4, Examples of retrieved images. In each case, the first image was the query.

5 Concluding Remarks

In this paper, we have showed the intrinsic relationship between appearance-based vision and VQ based image indexing techniques. We extended our earlier single resolution VQ based indexing method to multiresolution and introduced the idea of appearance indexing. Appearance indexing amounts to compile statistics of multiresolution appearance vectors approximated by the appearance prototypes, which in turn characterizes the contents of the image. We have successfully applied the method to content-based image retrieval and showed that it outperformed previous techniques.

Acknowledgements

This work was partially supported by a **Leverhulme Trust** Study Abroad Fellowship. The author is grateful to Professor Mark Drew, Department of Computing, Simon Fraser University, Canada, for hosting the Fellowship.

References

1. Poirson and B. Wandell, “Appearance of colored patterns: pattern-color separability”, J. Opt. Soc. Am. A, vol. 10, pp. 2458 – 2470, 1993
2. P. K. Kaiser and R. M. Boynton, Human Color Vision, Optical Society of America, Washington DC, 1996
3. W. Pratt, Digital Image Processing, Wiley, New York, 1978
4. CCIR, “Encoding parameters of digital television for studios”, CCIR Recommendation 601-2, Geneva, 1990
5. J. Huang et al, “Spatial color indexing and applications”, International Journal of Computer Vision, pp. 245 - 268, 1999
6. MPEG7 FCD, ISO/IEC JTC1/SC29/WG11, March 2001, Singapore
7. P. J. Burt and E. H. Adelson, “The Laplacian pyramid as a compact image code”, IEEE Trans. Commun., vol. 31, pp. 532 – 540, 1983
8. S. C. Ahalt, et al, “Competitive learning algorithms for vector quantization”, Neural Networks, vol. 3, pp. 277-290, 1990
9. J. Hertz, A. Krogh and R. Palmer, Introduction to the Theory of Neural Computation, Addison-Wesley, 1991
10. M. Swain and D. Ballard, "Color Indexing", International Journal of Computer Vision, Vol. 7, pp. 11-32, 1991
11. Schiele and J. L. Crowley, "Recognition without correspondence using multiresolution receptive field histogram", International Journal of Computer Vision, 36 (1), 31 -50, 2000
12. R. Rao and D. Ballard, “An active vision architecture based on iconic representations”, Artificial Intelligence, vol. 78, pp. 461-505, 1995
13. M. Turk and A. Pentland, “Eigenfaces for recognition”, Journal of Cognitive Neuroscience, vol. 3, pp. 71 – 86, 1991
14. H. Murase and S. K. Nayar, “Visual learning and recognition of 3D objects from appearance”, International Journal of Computer Vision, vol. 14, pp. 5-24, 1995
15. M. Kirby and L. Sirovich, “Low-dimensional procedures for the characterization of human faces”, Journal of Optical Society of America, vol. 4, pp. 519 –524, 1987
16. K. Nishino et al, “Eigen-texture mothod: Appearance compression and synthesis based on a 3D model”, IEEE Trans. PAMI, vol. 23, pp. 1257 –1265, 2001
17. K. Ohba et al, “Appearance-based visual learning and object recognition with illumination invariace”, Machine Vision and Applications, 12: 189 – 196, Springer-Verlag, 2000
18. I. Ulrich and I. Nourbakhsh, “Appearance-based obstacle detection with monocular color vision”, Proceedings of the AAAI National Conf. on AI, Austin, TX, July/August 2000
19. H. Schneiderman and T. Kanade, “A statistical method for 3D object detection applied to faces and cars”, IEEE CVPR 2000, pp. 746 – 751
20. B. Funt and G. Finlayson, “Color constant color indexing”, IEEE Reans PAMI, vol. 17, pp. 522-529, 1995
21. A. W. M. Smeulders et al, "Content-based image retrieval at the end of the early years", IEEE Trans PAMI, vol. 22, pp. 1349 - 1380, 2000
22. G. Qiu, "Pattern colour separable image coding for content based indexing", Proc. of IEEE International Workshop on Multimedia Signal Processing, Copenhagen, Denmark, September 13 -15, 1999 , pp. 407 – 412
23. G. Qiu, "Image indexing using a coloured pattern appearance model", Storage and Retrieval for Media Databases, 21-26 January 2001, San Jose, CA, USA
24. G Qiu, "Indexing chromatic and achromatic patterns for content-based colour image retrieval", Pattern Recognition, vol. 35, pp. 1675 – 1686, 2002
25. G. Lu and S. Teng, "A novel image retrieval technique based on vector quantization", Proc. Intl. Conf. on Computational Intelligence for Modelling, Control and Automation, 17-19 Feb. 1999, Viana, Austria, pp. 36-41.
26. Idris, F. and S. Panchanathan, “Image and video indexing using vector quantization”, Machine Vision and Applications, vol. 10, pp. 43-50. 1997
27. N. Vasconcelos and A. Lippman, “Feature representation for image retrieval: Beyond the color histogram”, IEEE Multimedia and Expo 2000
28. M. Jenkins and L. Harris (Eds.), Computational and Psychophysical mechanisms of visual coding, Cambridge University Press, 1997
29. A. Gersho, R. M. Gray, Vector quantization and signal compression, Kluwer Academic Publishers, Boston, 1992
30. E. Hadjidemetriou, M. Grossberg, and S. K. Nayar, “Spatial Information in Multiresolution Histograms”, Proceedings of IEEE 2001 Conference on Computer Vision and Pattern Recognition