(tiny)
ICPR 96

Object Location Using Complementary Color Features: Histogram and DCT

V V Vinod Hiroshi Murase

NTT Basic Research Labs, 3-1 Morinosato Wakamiya, Atsugi-shi, 243-01 Japan
E-mail: {vinod,murase}@apollo3.brl.ntt.jp

Abstract

Color constitutes an important cue for recognizing and locating objects in complex scenes. Most of the existing techniques using color employ only color histograms for object recognition and/or location. Color histograms are stable but not accurate. In this paper we study the complementary nature of color histogram and DCT coefficients with respect to accuracy and stability and develop a combined method using both histograms and DCT coefficients. The methods are experimentally evaluated. The combined method has higher stability and accuracy than using either feature alone.

1 Introduction

Object detection and location has several applications such as target selection, tracking, retrieval by image content, etc. The common approaches employed for this task are matching geometric features [1, 6] and template matching [13, 15, 18]. Color provides a powerful cue for image matching and has been proposed for content based image retrieval [5, 7, 11, 14, 16, 21]. Most of these systems employ color histogram matching. Swain and Ballard [17] proposed Histogram Intersection and Histogram Backprojection for object recognition and location. Local histogram intersection has been proposed in [4] for object location. We have developed Focussed Color Intersection with active search [19, 20] for efficiently recognizing and locating objects irrespective of size. Since histogram ignores the spatial distribution of colors, these methods are stable against template misalignments, changes in orientation, occlusion etc. However, this also leads to inaccurate locations. Combining color histogram and a feature taking into account the spatial distribution of color could give both stability and accuracy.

The discrete cosine transform (DCT) [3] used in image compression methods such as JPEG and MPEG [9] presents one such feature. DCT representation has been used for efficient eigenvalue decompositions in [12] and for scene cut detection in [8]. Since the spatial distribution of colors is taken into account, DCT will be a more accurate representation than histogram. However, it will not have the stability of histograms. For example, template misalignments could result in large location errors. Two stage template matching methods [18, 10] have been proposed for reducing computational cost. In this paper we propose two stage methods which combine the complementary nature of histogram and DCT for higher accuracy and stability.

The performance of methods using only color histogram or DCT alone are compared in section 2. Algorithms combining both histogram and DCT are presented in section 3. Experimental results are given in section 4 and conclusions in 5.

2 Single Feature Methods

The problem of recognizing and locating a model's instance in an image is essentially to determine *the location in the image at which the model, at some scale, matches best with a part of the image.* For best results the algorithm has to be *accurate* and *stable.* In this section we evaluate the accuracy and stability of histogram backprojection and focussed color intersection which uses color histograms and focussed DCT matching which uses DCT coefficient vectors.

The following abbreviations shall be used.

BP	Histogram Backprojection
FCI	Focussed Color Intersection
FDCT	Focussed DCT Matching
BP+FDCT	Combination of BP and FDCT
FCI+FDCT	Combination of FCI and FDCT

Also by a *model* we shall mean the reference image of the model.

2.1 Histogram Backprojection (BP)

Histogram backprojection [17] assigns a confidence value $C(x,y)$ to each location in the image as

$$C(x,y) = \frac{H_i^M}{H_i^I}$$

where H^M is the histogram of the model, H^I is the histogram of the image and pixel $\mathbf{p}_{xy}$ maps to histogram bin i. The peak of the smoothed confidence values gives the location of the object.

2.2 Focussed Color Intersection (FCI)

Focussed Color Intersection evaluates the match between the model and parts of the scene taking into account all possible sizes and locations [19]. This is done by scanning the scene at different resolutions with a fixed size window. Let the given image be of $N \times N$ pixels. Let $\mathbf{p}_{xy}^k$ denote pixels belonging to the image resized to $k \times k$ pixels. Consider scanning the images using a $w \times w$ pixels window shifted by s pixels along one direction at time. Then the set of focus regions are given by

$$\begin{aligned} \{R_{ij}^k\} \quad \text{where } k &= w, w+\Delta k, w+2\Delta k, \ldots, N \\ i &= 0, \ldots, \frac{k-w}{s}, \quad j = 0, \ldots, \frac{k-w}{s} \\ R_{ij}^k &= \{\mathbf{p}_{xy}^k \text{ such that } si \le x < si+w, \\ &\quad sj \le y < sj+w\} \end{aligned}$$

We shall refer to Δk as the *size parameter* and to s as the *shift parameter*. The histogram intersection [17] of a focus region R and a model M with normalized histograms H^R and H^M respectively is defined as

$$\text{Inter}_R = \sum_{i=1}^{b} \min\{H_i^R, H_i^M\}$$

where b is the number of histogram bins. A confidence value $C(x,y)$ is assigned to each (x,y) in the given image as

$$C(x,y) = \max_{R \in \mathsf{R}(x,y)} \text{Inter}_R$$

where $\mathsf{R}(x,y)$ is the set of all focus regions whose center points corresponds to location (x,y) in the $N \times N$ image. Locations not corresponding to the center of any focus region are assigned a confidence value of 0. The location with the highest confidence gives the object's location in the image.

2.3 Focussed DCT Matching

In focussed DCT matching the Euclidean distance between the normalized DCT coefficient vectors of the model and the focus regions is used as a measure of match confidence. Let Dist_R denote the Euclidean distance for focus region R. And let $\mathsf{R}(x,y)$ denote the set of all focus regions whose center corresponds to (x,y) in the given image. Then a confidence value is assigned to a location (x,y) as

$$C_M(x,y) = 2 - \max_{R \in \mathsf{R}(x,y)} \text{Dist}_R$$

The locations not corresponding to the center of any focus region are assigned a confidence measure of zero. The point (x,y) with the highest confidence value gives the location of the object. The normalized DCT coefficient vectors are constructed as follows.

The DCT of an $N \times N$ 2-dimensional signal $\mathbf{p}_{xy}$ may be defined as

$$\mathbf{t}_{uv} = \frac{2c_u c_v}{N} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} \mathbf{p}_{xy} \cos\frac{(2x+1)u\pi}{2N} \cos\frac{(2y+1)v\pi}{2N}$$
$$\text{for } u = 0, \ldots, N,\ v = 0, \ldots, N$$
$$\text{and } c_u, c_v = \tfrac{1}{\sqrt{2}} \text{ for } u, v = 0 \text{ and } 1 \text{ otherwise}$$

where x, y are spatial coordinates in the signal domain, u, v are coordinates in the transform domain and $\mathbf{t}_{uv}$ are the DCT coefficients. The coefficient $\mathbf{t}_{00}$ denotes the static component of the signal. $\mathbf{t}_{01}$ and $\mathbf{t}_{10}$ denote the lowest frequencies along y and x directions respectively. Increasing indices represent signal components with higher frequencies.

Following standard image compression algorithms, we adopt a block size of 8×8 for computing the DCT coefficients. This requires that the window size w and the model image size be multiples of 8. For an 8×8 block we will get 64 coefficients for each color component. Since the lower frequency components carry most of the relevant information it would be sufficient to construct the feature vector using a number of lower frequency coefficients out of the total 64. The sets $\{\mathbf{t}_{00}\}, \{\mathbf{t}_{01}, \mathbf{t}_{10}\}, \{\mathbf{t}_{20}, \mathbf{t}_{11}, \mathbf{t}_{02}\}, \ldots, \{\mathbf{t}_{67}, \mathbf{t}_{76}\}, \{\mathbf{t}_{77}\}$ constitute the sets of coefficients in order of increasing frequency. If n is the number of lowest frequency coefficient sets chosen then the DCT feature vector is constituted by $\mathbf{t}_{uv}$ such that $u+v < n$. The coefficient selection maps for $n = 1, 4$ and 6 are shown below. A '1' indicates that the coefficient in that position is selected and '0' indicates that it is not selected.

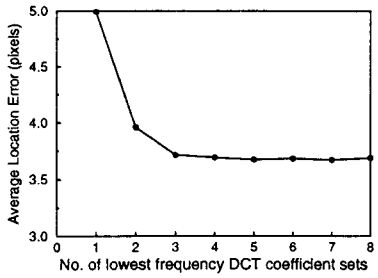

Figure 1. Average location error vs. number of lowest frequency DCT coefficient sets

$n=1$	$n=4$	$n=6$
10000000	11110000	11111100
00000000	11100000	11111000
00000000	11000000	11110000
00000000	10000000	11100000
00000000	00000000	11000000
00000000	00000000	10000000
00000000	00000000	00000000
00000000	00000000	00000000

The average location error for n ranging from 1 to 8, with the size parameter Δk and shift parameter s fixed at 4 pixels are shown in figure 1. Figure 1 indicates that $n > 3$ would be a good choice and $n = 4$ was used in all other experiments. The DCT coefficient vector for a focus region (model) is constructed using 4 lowest frequency coefficients of all color components in each 8×8 block in the focus region (model). This vector is normalized (Euclidean norm = 1) and used as the feature for matching.

FDCT may be applied to JPEG and MPEG compressed data without full reconstruction using the method in [2]. The DCT feature vector of focus regions aligned with the 8×8 blocks will be readily available. For other focus regions, not aligned with the 8×8 blocks, the DCT coefficients can be derived using the method in [2]. However, for FDCT the DCT coefficients of the image at different resolutions are required. Moreover the complexity of the method given in [2] is not much different from performing full reconstruction and DCT computation. Therefore, in general, the gain in computational effort may not be large.

2.4 Comparative Results

The average location error obtained using BP, FCI and FDCT over 300 images, each of 128×128 pixels, are shown in figure 3. Two sample images and models are shown in figure 2. The location error was computed as the Euclidean distance between the manually determined actual location and the location determined by the algorithm.

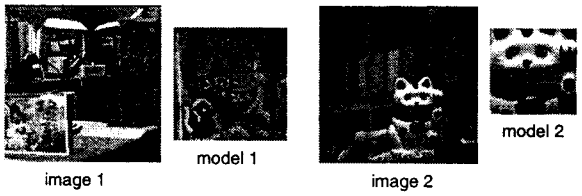

Figure 2. Two sample images and models

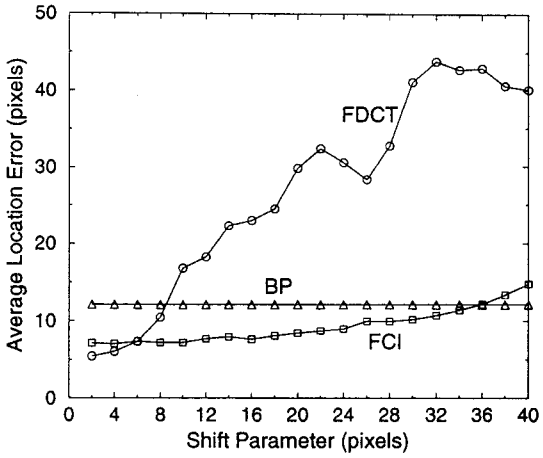

Figure 3. Comparison of algorithms BP (histogram backprojection), FCI (focussed color intersection) and FDCT (focussed DCT matching)

The RGB histogram with 4096 bins were used for BP and FCI. In BP, Gaussian smoothing with variance 6 and a kernel size of 32×32 was employed for smoothing the confidence values. Focus regions for FCI and FDCT were constructed with a 32×32 pixel scanning window and the size parameter was fixed at 4 pixels.

The results are plotted in figure 3 for s ranging from 2 to 40. When a low value of s is used, the image is densely scanned and an accurate algorithm will determine the correct location. At low values of s the image is scanned densely and the computational effort is hight. Higher s implies sparsely searching the image at low computational effort. An accurate algorithm will have low error at low s and a stable algorithm will degrade gracefully with increasing s. From the results, the stability and accuracy of the algorithms may be summarized as follows:

Algorithm	Accuracy	Stability
BP	low	medium
FCI	medium	high
FDCT	high	low

3 Combining Histogram and DCT

From section 2.4 we see that color histogram based methods are more stable than DCT coefficient matching and the latter is more accurate. Figure 4 shows the normalized confidence values associated to different locations in image 2 of figure 2 given model 2 of figure 2. The confidence values are quantized into 10 different levels. Lighter regions have higher confidence values and darker regions have lower confidence values.

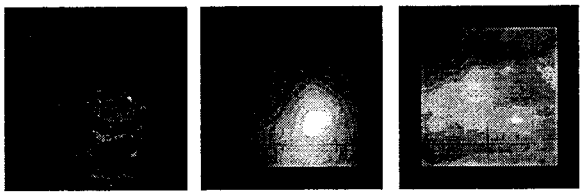

Figure 4. Confidence values given by algorithms BP (histogram backprojection), FCI (focussed color intersection) and FDCT (focussed DCT matching)

From the distribution of confidence values the following may be observed:

BP Pixels having same color as pixels in the model are given high confidence values. In general, the accuracy and stability would depend greatly on background pixels and object sizes.

FCI There is a single large region with highest confidence. There is one contiguous region for one confidence level which is surrounded by a region having the next lower confidence level. This indicates high stability.

FDCT There is a small region with highest confidence level very near the correct location of the object. For the next confidence level there are three disjoint, widely separated regions. If the highest confidence region is missed (say due to template misalignment) then a totally false location may be identified.

The unstable nature of FDCT can be avoided if FDCT is restricted to an area near the actual location. Both BP and FCI associates high confidence values to areas near the actual location. The combined methods restricts FDCT to such areas and thereby provide higher stability and accuracy.

3.1 Two Stage Methods

The combined methods involves two steps - candidate selection using histogram and matching candidates against the model using focussed DCT matching. The scheme is shown in in figure 5.

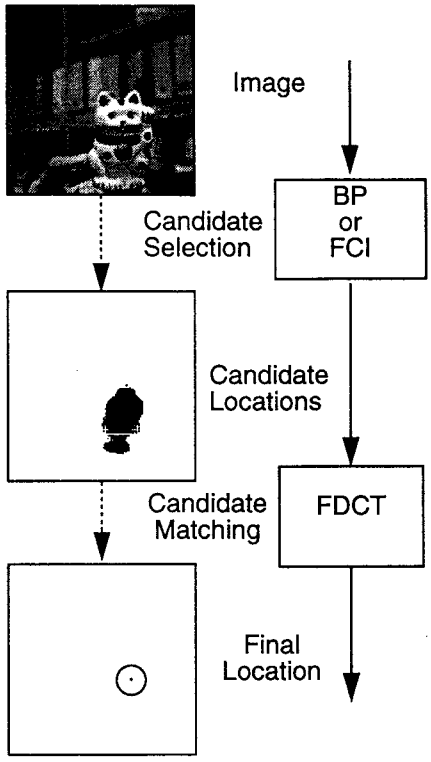

Figure 5. The combined methods

We consider candidate selection using either BP or FCI in two algorithms BP+FDCT and FCI+FDCT. Using algorithm BP the confidence values are obtained as given in section 2.1. For candidate selection using FCI the confidence values of different locations in the image are defined as follows.

$$C(x,y) = \max_{R \in \mathsf{R}(x,y)} \mathrm{Inter}_R$$

where $\mathsf{R}(x,y)$ is the set of all focus regions covering location (x,y) in the scene. This definition associates a confidence value for all locations in the image. Since accurate location is not the objective, FCI for candidate selection can be applied with large values of size parameter Δk and shift parameter s. This will result in a flat distribution of the confidence values. Hence, Gaussian smoothing with kernel size 16×16 and variance 3.0 was used for smoothing the confidence values obtained by FCI. The confidence values obtained by BP or FCI are smoothed and normalized such that

	θ	N_{DCT}	*Error*
BP+DCT	0.7	930	5.6
	0.6	1431	4.3
	0.4	2263	3.5
FCI+DCT	0.999	849	4.0
	0.99	1367	3.5
	0.95	2502	3.4

Table 1. Results of BP+DCT and FCI+DCT for different candidate selection thresholds

the highest confidence value is 1.0. All locations with the smoothed and normalized confidence value greater than a threshold θ are selected as candidate locations.

In the second step the FDCT algorithm is applied to the candidate locations to determine the final location. FDCT is applied with small values of s for obtaining accurate results. Since the candidate locations will be few in number the number of focus regions matched and consequently the computational effort will be low.

3.2 Active Search for Candidate Selection

Active search [20] have been proposed for efficiently determining the best matching focus region without evaluating the histogram intersection at all focus regions. The search is directed by upper bound estimates. The algorithm decides the next focus region to be matched based on the histogram intersections of focus regions matched till then. The search concentrates on focus regions having higher histogram intersections. By this process, substantial gain in computational effort is achieved without loss of accuracy.

Active search may be applied for candidate selection with an upper bound cutoff equal to the threshold θ times the highest histogram intersection value encountered by the algorithm. It is clear that this will ensure that all candidates are identified. At the same time large number of focus regions will not be matched reducing the computational effort for candidate selection.

4 Experimental Results

The images and models used in the present experiments are the same as those used for experiments in section 2.4. We shall denote the average location error by *Error* and the number of DCT coefficient matches and histogram intersections evaluated by N_{DCT} and N_{HI} respectively.

In table 1 we show the results obtained for different values of the candidate selection thresholds with shift

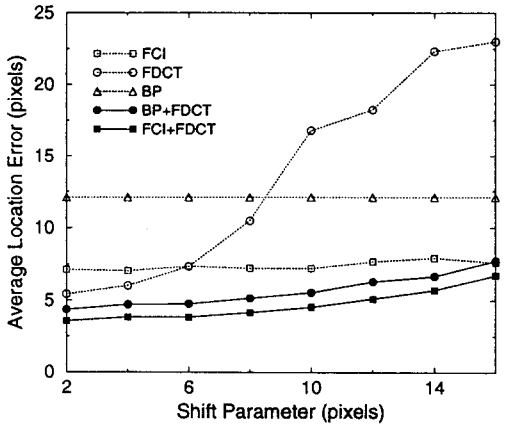

Figure 6. Comparison of algorithms BP, FCI, FDCT, BP+FDCT and FCI+FDCT

parameter $s = 2$ and size parameter $\Delta k = 4$. From the results we observe that, for similar values of N_{DCT} FCI+FDCT has lower error than BP+FDCT. Candidate selection thresholds of 0.6 for BP+FDCT and 0.99 for FCI+FDCT are adopted in other experiments. It may be noted that, for these thresholds, the number of focus regions at which DCT matching is done are approximately same for both algorithms.

The results obtained using the combined algorithms for different values of the shift parameter are plotted given in figure 6. The candidate selection step in FCI+FDCT employed size parameter $\Delta k = 8$ and shift parameter $s = 16$. In the second step, FDCT was applied to the candidate locations with size parameter $\Delta k = 4$, for both BP+FDCT and FCI+FDCT. The shift parameter used was varied from 2 to 16 in steps of 2. The average location errors for BP, FCI and FDCT are also reproduced from figure 3 for comparison. It may be observed from figure 6 that the average location error and the increase in error with increasing s are both less for the combined algorithms. Thus BP+FDCT and FCI+FDCT are more stable and accurate than BP, FCI or FDCT.

In table 2 we compare the computational effort ($N_{DCT} + N_{HI}$) and the average location error of all the five methods. It may be observed that for low values of s active search is very efficient and only about 15% of the total number of focus regions are matched by FCI. For candidate selection with shift parameter 16, FCI matched only less than 50% out of the total 231 focus regions.

The results in table 2 indicate the higher accuracy and stability of the combined methods at low compu-

method	shift parameter 2		shift parameter 10	
	N_{HI}+ N_{DCT}	*Error*	N_{HI} N_{DCT}	*Error*
BP	-	**12.0**	-	12.0
FCI	3116	**7.1**	612	7.2
FDCT	20825	**5.4**	935	16.0
BP+DCT	1431	**4.3**	59	5.5
FCI+DCT	1475	**3.5**	165	4.5

Table 2. Results of combined method for two values of shift parameter

tational effort. For $s = 2$ BP+FDCT and FCI+FDCT achieve higher accuracy at less than 10% of the effort of FDCT. At higher values of s also the combined algorithms have higher accuracy and lower computational effort. Thus, BP+DCT and FCI+DCT are ideally suited for fast, stable and accurate object recognition and location. FCI+FDCT provides higher accuracy for small amount of extra computational effort.

5 Conclusion

We have studied the complementary nature of color histogram and DCT coefficients for locating colored objects. Two stage methods combining histogram and DCT have been proposed and shown to be more accurate, stable and computationally efficient. Active search for improving the efficiency of candidate selection using focussed color intersection has also been discussed.

Acknowledgements The authors wish to thank Dr. T. Ikegami, Dr. K.Ishii, Dr. N.Hagita and Dr. S.Naito of NTT Basic Research Labs for their help and encouragement in conducting this research.

References

[1] N. A. Ayache and O. D. Fougeras. HYPER: A new approach for the recognition and location of two-dimensional objects. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8:44–54, 1986.

[2] S.-F. Chang and D. G. Messerschmitt. A new approach to decoding and compositing motion-compensated dct-based images. In *Proc. of IEEE Int. Conf. Acoust., Speech Signal Proc.*, pages V–421–V–424, 1993.

[3] W. H. Chen and H. Smith. Adaptive coding of monochrome and color images. *IEEE Trans. Communication*, COM-25:1285–1292, 1977.

[4] F. Ennesser and G. Medioni. Finding waldo, or focus of attention using local color information. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17:805–809, 1995.

[5] M. Flickner et al. Query by image and video content : The QBIC system. *IEEE Computer*, 28(9):23–32, Sept. 1995.

[6] W. E. L. Grimson. *Object Recognition by Computer: The Role of Geometric Constraints.* The MIT Press, 1990.

[7] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(7):729–736, 1995.

[8] E. Iwanari and Y. Ariki. Scene clustering and cut detection in moving images by dct components. *Technical Report of IEICE*, PRU 93-119:23–30, 1994.

[9] D. Legall. MPEG- a video compression standard for multimedia applications. *Communications of the ACM*, 34(4):46–58, 1991.

[10] A. Margalit and A. Rosenfeld. Using feature probabilities to reduce the expected computational cost of template matching. *Computer Vision Graphics and Image Processing*, 52:110–123, 1990.

[11] B. M. Mehtre, M. S. Kankanhalli, A. D. Narasimhalu, and G. C. Man. Color matching for image retrieval. *Pattern Recognition Letters*, 16:325–331, 1995.

[12] H. Murase and M. Lindenbaum. Partial eigenvalue decomposition of large images using spatio temporal adaptive method. *IEEE Trans. Image Proc.*, 4:620–629, 1995.

[13] H. Murase and S. K. Nayar. Image spotting of 3d objects using parametric eigenspace representation. In *Proceedings of the 9th Scandinavian Conference on Image Analysis*, June 1995.

[14] A. Nagasaka and Y. Tanaka. *Automatic Video Indexing and Full-Video Search for Object Appearances*, pages 113–127. Elsevier Science Publishers, B.V., 1992.

[15] A. Rosenfeld and G. J. Vanderburg. Coarse-fine template matching. *IEEE Trans. System, Man and Cybernetics*, SMC-7:104–107, 1977.

[16] R. Schettini. Multicolored object recognition and location. *Pattern Recognition Letters*, 15:1089–1097, 1994.

[17] M. J. Swain and D. H. Ballarad. Indexing via color histograms. In *Proc. Image Understanding Workshop*, pages 623–630, 1990.

[18] G. J. Vanderburg and A. Rosenfeld. Two stage template matching. *IEEE Trans. Comput.*, C-26:384–393, 1977.

[19] V. V. Vinod and H. Murase. Focussed color intersection for object extraction from cluttered scenes. In *Proc. of Vision Interface*, May 1996.

[20] V. V. Vinod, H. Murase, and C. Hashizume. Focussed color intersection with efficient searching for object detection and image retrieval. In *Proc. of IEEE Conference on Multimedia Computing Systems*, June 1996.

[21] J. K. Wu, A. D. Narasimhalu, B. M. Mehtre, C. P. Lam, and Y. J. Gao. CORE: a content-based retrrieval engine for multimedia information systems. *ACM Multimedia Systems*, 3(1):25–41, 1995.