Colour Image Retrieval and Object Recognition Using the Multimodal Neighbourhood Signature

Jiri (George) Matas[1,2], Dimitri Koubaroulis[1], and Josef Kittler[1]

[1] CVSSP, University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom
{D.Koubaroulis, G.Matas}@ee.surrey.ac.uk
[2] CMP, Czech Technical University, 121 35 Prague, Czech Republic

Abstract. A novel approach to colour-based object recognition and image retrieval -the multimodal neighbourhood signature- is proposed. Object appearance is represented by colour-based features computed from image neighbourhoods with multi-modal colour density function. Stable invariants are derived from modes of the density function that are robustly located by the mean shift algorithm. The problem of extracting local invariant colour features is addressed directly, without a need for prior segmentation or edge detection. The signature is concise – an image is typically represented by a few hundred bytes, a few thousands for very complex scenes.
The algorithm's performance is first tested on a region-based image retrieval task achieving a good (92%) hit rate at a speed of 600 image comparisons per second. The method is shown to operate successfully under changing illumination, viewpoint and object pose, as well as non-rigid object deformation, partial occlusion and the presence of background clutter dominating the scene. The performance of the multimodal neighbourhood signature method is also evaluated on a standard colour object recognition task using a publicly available dataset. Very good recognition performance (average match percentile 99.5%) was achieved in real time (average 0.28 seconds for recognising a single image) which compares favourably with results reported in the literature.

1 Introduction

Colour-based image and video retrieval has many applications and acceptable results have been demonstrated by many research and commercial systems during the last decade [22]. Very often, applications require retrieval of images where the query object or region cover only a fractional part of the database image, a task essentially identical to appearance-based object recognition with unconstrained background. Retrieval and recognition based on object colours must take into account the factors that influence formation of colour images: viewing geometry, illumination conditions, sensor spectral sensitivities and surface reflectances. In many applications, illumination colour, intensity as well as view point and background may change. Moreover, partial occlusion and deformation of non-rigid objects must also be taken into consideration. Consequently, invariance or at least robustness to these diverse factors is highly desirable.

Most current colour based retrieval systems utilise various versions of the colour histogram [24] which has proven useful for describing the colour content of the whole image. However, histogram matching cannot be directly applied to the problem of recognising objects that cover only a fraction of the scene. Moreover, histograms are not invariant to varying illumination and not generally robust to background changes. Applying colour constancy methods to achieve illumination invariance for histogram methods is possible but colour constancy itself poses a number of challenging problems [8]. Other methods addressing image (as opposed to object) similarity are those using wavelets [12] and moments of the colour distribution [11, 18].

Finally, graph representations of colour content (like the colour adjacency graph [15] and its extension to a hybrid graph [21]) have provided good recognition for scenes with fairly simple colour structure.

Departing from global methods, localised invariant features have been proposed in order to gain robustness to background changes, partial occlusion and varying illumination conditions. Histograms of colour ratios computed locally from pairs of neighbouring pixels for every image pixel [9] or across detected edges [10] have been used. However, both methods are limited due to the global nature of histogram representation. In the same spirit, invariant ratio features have been extracted from nearby pixels across boundaries of segmented regions for object recognition [20, 19]. Absolute colour features have been extracted from segmented regions in [23, 17]. However, reliable image segmentation is arguably a notoriously difficult task [22, 19]. Other methods split the image into regions from where local colour features are computed. For example, the FOCUS system [5] constructs a graph of the modes of the colour distribution from every image block. However, not only extracting features from every image neighbourhood is inefficient, but also the features used do not account for illumination change. In addition, use of graph matching for image retrieval has often been criticised due to its relatively high complexity.

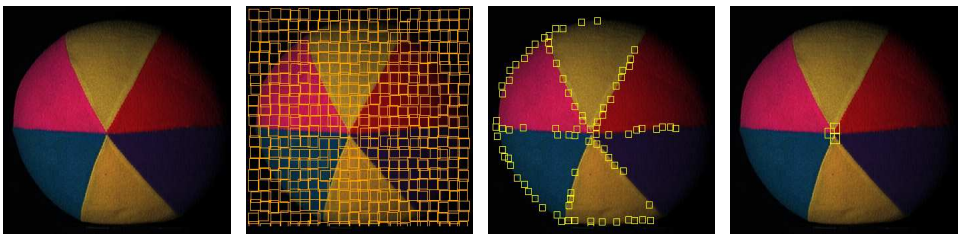
We propose a method to address the colour indexing task by computing colour features from local image neighbourhoods with multimodal colour probability density function. First, we detect multimodal neighbourhoods in the image using a robust mode estimator, the *mean shift algorithm* [7]. From the mode colours we are then able to compute a number of local invariant features depending on the adopted model of colour change. Under different assumptions, the resulting multimodal neighbourhood signatures (MNS) consist of colour ratios, chromaticities, raw colour values or combinations of the above. Our method improves on previous ones by

- creating a signature which concisely represents the colour content of the image by stable measurements computed from neighbourhoods with informative colour structure. Neither prior segmentation nor edge detection is needed.
- computing invariant features from robustly filtered colour values representing local colour content
- effectively using the constraints about the illumination change model thus resulting in a flexible colour signature
- applying signature instead of histogram matching to identify and localise the query object in the database images

The advantages of computing features from detected multimodal neighbourhoods are discussed in the next section. The algorithmic details of the approach and the implemented algorithm is described in section 3. Section 4 presents details about the experimental setup and the results obtained are presented in section 5. Section 6 concludes the paper.

2 The MNS Approach

Consider an image region consisting of a small compact set of pixels. The shape of the region is not critical for our application. For convenience, we use regions defined as neighbourhoods around a central point. Depending on the number of modes of the probability distribution of the colour values, we characterise such regions as unimodal or, for more than one mode, multimodal neighbourhoods. Clearly, for unimodal neighbourhoods no illumination invariant features can be computed. We therefore focus on detected multimodal neighbourhoods. In particular, multimodal neighbourhoods with more than two modes provide good characterisation of objects



(a) (b) (c) (d)

Fig. 1. Multimodal neighbourhood detection: (a) original image (b) randomised grid (c) detected bimodal neighbourhoods (d) detected trimodal neighbourhoods

like the ball in Fig 1(d) and can result in efficient recognition on the basis of only few features.

The advantages of extracting colour information from multimodal neighbourhoods are many-fold. Local processing is robust to partial occlusion and deformation of non-rigid objects. Data reduction is achieved by extracting features only from a subset of all image neighbourhoods. Moreover, a rich description of colour content is obtained since a single colour patch can contribute to more than one neighbourhood feature computation. The computation time needed to create the colour signature is small since the most common neighbourhood type - unimodal neighbourhoods - are ignored after being detected very efficiently. Furthermore, illumination invariant features can be computed from the mode values to account for varying illumination conditions even within the same image. Regarding retrieval, region-based queries are efficiently handled and localisation of the query instance in the database images is possible. Finally, the proposed representation allows the users to select exactly the local features they are interested in from the set of the detected multimodal neighbourhoods of the query image.

Computing colour invariants from detected multimodal neighbourhoods has certain advantages with respect to extracting features across detected edges. Most edge detection methods require an intensity gradient and a locally linear boundary. They often perform poorly at corners, junctions and regions with colour texture - exactly in those regions, where colour information can be highly discriminative. In addition, the multimodal neighbourhood approach directly formulates the problem of extracting colour features.

3 The Algorithm

3.1 Computing the MNS Signature

The image plane is covered by a set of overlapping small compact regions. In the current implementation, rectangular neighbourhoods with dimensions (b_x, b_y) were chosen. Compact regions of arbitrary shape - or even non-contiguous compact sets of pixels - could have been used. Rectangular neighbourhoods were selected since they facilitate simple and fast processing of the data. To avoid aliasing each rectangle is perturbed with a displacement with uniform distribution in the range $[0, b_x/2)$, $[0, b_y/2)$, Fig. 1(b). To improve coverage of an image (or image region), more than one randomised grids can be used, slightly perturbed from each other.

For every neighbourhood defined by such randomised grids, the modes of the colour distribution are computed with the mean shift algorithm described below.

Modes with relatively small support are discarded as they usually represent noisy information. The neighbourhoods are then categorised according to their modality as unimodal, bimodal, trimodal etc. (e.g. see Fig. 1)

For the computation of the colour signature only multimodal neighbourhoods are considered. For every pair of mode colours m_i and m_j in each neighbourhood, we construct a vector $v = (m_i, m_j)$ in a joint 6-dimensional domain denoted RGB^2. In order to create an efficient image descriptor, we cluster the computed colour pairs in the RGB^2 space and a representative vector for each cluster is stored. The colour signature we propose consists of the modes of the distribution in the RGB^2 space. For the clustering, the mean shift algorithm is applied once more to establish the local maxima. The computed signature consists of a number of RGB^2 vectors depending on the colour complexity of the scene. The resulting structure is, generally, very concise and flexible.

Note that for the computation of the signature no assumption about the colour change model was needed. The parameters controlling mode seeking, that is the kernel width and the neighbourhood size are dependent on the database images; the former being related to the amount of filtering (smoothing) associated with the mean shift and the latter depending on the scale of the scene. A multiscale extension of the algorithm, though relatively straightforward to implement (e.g. by applying the MNS computation to an image pyramid), has not yet been tested.

3.2 Computation of Neighbourhood Modality with the Mean Shift Algorithm

To establish the location of a mode of the colour density function the mean shift algorithm is applied in the RGB domain. The general kernel-based estimate of a true multivariate density function $f(\bar{x})$ at a point $\bar{x}_0$ in a d-dimensional data space is given by

$$\hat{f}(\bar{x}_0) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\bar{x}_i - \bar{x}_0}{h}\right) \tag{1}$$

where $\bar{x}_i,\ i = 1..n$ are the sample data points and K is the kernel function with kernel width h. In this work, we are not interested in the value of the density function at the point $\bar{x}_0$ but rather in the location of its maxima locations in the data space. A simple and efficient algorithm for locating the maximum density points was proposed by Fukunaga [7] when the kernel function in (1) is the Epanechnikov kernel

$$K_E(\bar{x}) = \begin{cases} \frac{1}{2} c_d^{-1}(d+2)(1 - \bar{x}^T\bar{x}) & \text{if } \bar{x}^T\bar{x} < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where c_d is the volume of the unit d-dimensional sphere and $\bar{x}$ are the data points. The kernel has been shown to be robust to outliers and optimum in the sense of having minimum integrated square error in comparison with other kernels [3].

The mechanism of the mean shift algorithm consists of iteratively shifting the kernel to the average of the data points within by the mean difference vector

$$M_h(\bar{x}) = \frac{1}{n_{\bar{x}}} \sum_{\bar{x}_i \in S_h(\bar{x})} (\bar{x}_i - \bar{x}) = \frac{h^2}{d+2} \frac{\hat{\nabla} f(\bar{x})}{\hat{f}(\bar{x})} \tag{3}$$

where $n_{\bar{x}}$ is the number of data points inside the hypersphere S of radius h centred at $\bar{x}$. Equation 3 is an estimate of the normalised gradient of the density function $f(\bar{x})$ in the d-dimensional spac. As shown in [7], translation of the kernel centre towards the direction of the mean difference vector is equivalent to a gradient ascent to the local mode of the distribution. Convergence to the closest mode is guaranteed [4].

(a) (b)

Fig. 2. Robust filtering using the mean shift algorithm: (a) original image (b) filtered image (every neighbourhood pixel replaced by the mode of the density function it converged to)

Due to the non-linearity of the kernel, the filtering preserves discontinuities, details and retains local image structure. This is particularly important for images containing small objects like the swimmer's cap in Fig. 2. The speed of the algorithm was tested experimentally, and convergence was very fast (typically 4-5 iterations for complex data). Due to its advantageous properties the mean shift algorithm has been used in the past for image segmentation [4] and face tracking. For the MNS method, a computationally simple algorithm was implemented (see [16] for an efficient implementation).

Replacing each pixel in the neighbourhood with the mode it converged to results in a filtered image like the one in Fig. 2(b). The filtered image is produced by replacing the value of each pixel p_j , $j = 1..n$ of a neighbourhood with the closest mode m_j of the 3-dimensional colour density function using an iterative procedure:

For each $j = 1..n$

1. Initialise $i = 0$ and set the current mode estimate m_j^0 to the value of the pixel p_j
2. Update the mode estimate $m^{i+1} = \frac{1}{n_i} \sum_{\bar{x}_k \in S_3(m^i)} \bar{x}_k, \quad i \leftarrow i + 1$ until convergence i.e. until $m^{i+1} - m^i < \epsilon$
3. Replace the value of pixel p_j with the value of the local mode m_j it converged to.

3.3 Computing Invariant Features from Multimodal Neighbourhoods

From the multimodal neighbourhood signature, a number of invariant features can be computed. For the ease of exposition we will describe feature extraction from bimodal neighbourhoods which are the simplest multimodal ones.

Consider a local image patch with two adjacent surfaces i and j. According to the monochromatic model of surface reflectance [15, 10] the two estimated mode colours will be given by

$$\begin{aligned} r_i &= (R_i, G_i, B_i) = s_i^k g_i c_i^k \\ r_j &= (R_j, G_j, B_j) = s_j^k g_j c_j^k \quad , \quad k = R, G, B \end{aligned}$$

where s_i^k is the illumination factor, g_i is the geometric factor and c_i^ks the k-th sensor response to the surface reflectance of patch i under white light(surface colour).

Besides modelling the effects of change in viewpoint and object pose, the geometric factor g of the monochromatic model encompasses all factors that have the same effect on each colour channel, e.g. change of aperture or camera gain and change in illumination intensity. Coefficients s_i^k represent factors that effect individual colour channels, e.g. the change of illumination colour in the diagonal colour constancy model described below.

A different image of the same surface colour pair under different light and object pose would change recorded colours to

$$\begin{aligned} r'_i &= (R'_i, G'_i, B'_i) = s'^k_i g'_i c^k_i \\ r'_j &= (R'_j, G'_j, B'_j) = s'^k_j g'_j c^k_j \quad , \quad k = R, G, B \end{aligned}$$

respectively.

Assuming constant illumination for both the database and the query scenes the colour change model for $s'^k_c = s^k_c$, $g'_c = g_c$, $c = i, j$ and $g_i = g_j$ becomes $r'_i = r_i$ and $r'_j = r_j$ and the simplest invariant colour features appear to be the mode colour values in the RGB^2 space

$$f_{sg} = (R_i, G_i, B_i, R_j, G_j, B_j)$$

When $s'^k_c = s^k_c$, $c = i, j$, $\frac{g_i}{g'_i} = \frac{g_j}{g'_j}$ but $g_i \neq g_j$, orientation change is assumed to be same for both surfaces under constant light. Colour change is modelled by

$$r'_i = \frac{g'_i}{g_i} r_i \text{ and } r'_j = \frac{g'_j}{g_j} r_j$$

In this case, various 5-dimensional features can be constructed from the mode chromaticities

$$x_k = \frac{R_k}{R_k + G_k + B_k}, \quad y_k = \frac{G_k}{R_k + G_k + B_k} \quad k = i, j$$

and rational features. For example the 2 mode chromaticities and an intensity ratio produce the 5D feature vector

$$f_g = (x_i, y_i, I_{ij}, x_j, y_j) \quad \text{where} \quad I_{ij} = (\frac{R_i + G_i + B_i}{R_j + G_j + B_j})$$

proposed in [15].

When $s'^k_c = s^k_c$, $c = i, j$ but $g_i \neq g_j$ and $\frac{g_i}{g'_i} \neq \frac{g_j}{g'_j}$, varying illumination intensity is assumed due to surface discontinuities and orientation changes. Colour change is modelled as before and the chromaticities of the estimated mode colours are simple invariant 4-dimensional features

$$f_{gd} = (x_i, y_i, x_j, y_j)$$

The assumption of constant illumination within the same scene is violated in most natural scenes. However, it is realistic to assume constant illumination colour in local image neighbourhoods. The diagonal model of illumination change has been shown plausible when camera sensors are sufficiently narrow-band filters [6]. According to this model, illumination change is modelled by an independent scaling of the colour channels by a different constant i.e. $s'^k_c = d_k s^k_c$, $c = i, j$, $d_k \in \mathbb{R}$. It is easy to show that (assuming diagonal illumination change) the ratio of colours between two neighbouring surfaces with different colours is invariant to lighting changes [14, 9]. Nevertheless, for the assumption to hold, the two neighbouring surfaces must

have the same orientation i.e. $g_i = g_j$. Invariant features can be computed from the 3 colour channel ratios of the mode RGB values

$$f_c = \left(\frac{R_i}{R_j}, \frac{G_i}{G_j}, \frac{B_i}{B_j} \right)$$

In the most general situation, where orientation is different for the two surfaces $g_i \neq g_j$ and $\frac{g_i}{g_i'} \neq \frac{g_j}{g_j'}$, the 2-dimensional cross-ratio vectors

$$f_{cgd} = \left(\frac{R_i G_j}{G_i R_j}, \frac{G_i B_j}{G_j B_i} \right)$$

are invariant under the diagonal model as shown in [10].

Different invariants, but not necessarily independent, may be computed from a pair of RGB values (e.g. based on the hue-saturation colour model). We have not explored this issue. Invariants that could be obtained by exploiting higher order information from neighbourhoods with more than 2 modes have not been studied either.

3.4 Matching Multimodal Neighbourhood Signatures

A simple signature matching technique was applied to compute the dissimilarity between two MNS image signatures. The algorithm attempts to find a match for all model features assuming that the model signature contains only information about the object of interest. This assumption is realistic, since in object recognition applications a model database is typically built off-line in controlled conditions (e.g. with background allowing easy segmentation). In image retrieval applications, the query region is delineated by the user. Sometimes the full image is the object of interest and its MNS description is an appropriate model. However, if only part of the image is covered by the object of interest and the full image descriptor is stored as a model, a loss in recognition performance is likely.

On the other hand, test images may originate from scenes containing the model (query) object only as a fraction of the picture. The matching procedure is therefore asymmetric. A mismatch of a model feature is penalised whereas a mismatch of a test image feature is not. In other words the matching algorithm attempts to interpret the model signature as a distorted subset of the test image signature.

Let $I = 1..n$ and $J = 1..m$ be the indices of the model and test features respectively. We define a match association function $u(i) : I \rightarrow 0 \bigcup J$, $i \in I$, mapping each model feature i to the test feature it matched or to 0 if it did not match. Similarly, a test association function $v(j) : J \rightarrow 0 \bigcup I$, $j \in J$, maps a test feature to a model feature or to 0 in case of no match. A single threshold T_h defines the maximum allowed distance between two matching features. The matching problem, i.e the problem of uniquely associating each feature $s_i^M, i = 1..n$ of the model signature with a test feature $s_j^T, j = 1..m$ and the computation of a match score is resolved in the following 4 steps:

1. Set $u(i) = 0 \quad \text{and} \quad v(j) = 0 \quad \forall\, i, j$. From each signature s compute the invariant features f_i^M, f_j^T according to the colour change model dictated by the application.
2. Compute all pairwise distances $d_{ij} = d(f_i^M, f_j^T)$ between the model and test features.
3. Set $u(i) = j$, $v(j) = i \quad \text{if} \quad d_{ij} < d_{kl}$ and $d_{ij} < T_h \quad \forall\, k, l \quad \text{with} \quad u(k) = 0$ and $\quad v(l) = 0$.
4. Compute signature dissimilarity as

$$D(s^M, s^T) = \sum_{(\forall\, i\,:u(i)\neq 0)} d_{ij} + \sum_{(\forall\, i\,:u(i)=0)} T_h \tag{4}$$

Computing overall image similarity, the quality of the model features that matched is taken into account and the score is penalised for any unmatched model features. Note that features are allowed to match only once. In general, the more model features matched, the lower the $D(s^M, s^T)$ value and the more similar the compared images.

3.5 Computing Feature Distances

Let $v = (v_a, v_b)$ and $u = (u_a, u_b)$ be two vectors in the RGB^2 space. The adopted distance function is the sum of the square norms of the pairwise vector component differences

$$d_{RGB^2}(v, u) = min\left\{\|v_a - u_a\| + \|v_b - u_b\|, \|v_a - u_b\| + \|v_b - u_a\|\right\} \tag{5}$$

Taking the minimum distance between the original and component-wise inverted vectors is necessary because the order of the mode values in the joint vectors is not fixed.

Various distance functions can be defined for the chromaticity and rational features. We chose the same function (5) for measuring distance in the 4D joint chromaticity domain. A distance for 5D features was proposed in [15]. For matching relative features, a simple formula was devised

$$d_{frac}(p, q) = \frac{|a * d - b * c|}{\sqrt{a + b + c + d}} \tag{6}$$

where $p = \frac{a}{b}$ and $q = \frac{c}{d}$ are 1-dimensional fractions. The distance between the colour ratio between two RGB values p_i, p_j defined as

$$r_1 = \left(r_R^1, r_G^1, r_B^1\right) = \left(\frac{R_i^1}{R_j^1}, \frac{G_i^1}{G_j^1}, \frac{B_i^1}{B_j^1}\right)$$

and another ratio $r_2 = \left(r_R^2, r_G^2, r_B^2\right)$ between two other colours q_i and q_j is then

$$d_{rat}(r_1, r_2) = \frac{1}{3}\left(d_{frac}(r_R^1, r_R^2) + d_{frac}(r_G^1, r_G^2) + d_{frac}(r_B^1, r_B^2)\right) \tag{7}$$

The modification of d_{rat} to measure the distance between 2-dimensional cross-ratios is trivial, ignoring one colour channel in (7).

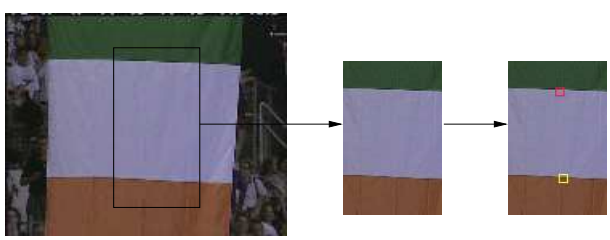


Fig. 3. Query selection and representative multimodal neighbourhoods

(a) (b) (c) (d)

Fig. 4. Sample target images demonstrating possible cases of: (a) background clutter, (b) non-rigid deformation, (c) illumination change and (d) object size

4 Data and Experimental Setup

4.1 Image Retrieval Experiment

We tested the suitability of the multimodal neighbourhood signature method for region-based image retrieval using a 30 minute video sequence of a BBC summary of the Atlanta Olympic games. The objective of the experiment was to retrieve frames that involved Irish events or athletes, therefore we searched for the presence of the Irish national colours in the image database.

In total, 145 frames were randomly chosen from the sequence yielding a database of very different images, taken both indoors and outdoors. Object pose, scale as well as illumination was arbitrary (Fig. 5). No image was removed from the original selection and no image preprocessing was applied. The size of each frame was 176×144 pixels. The query image was a rectangular region, a part of an Irish flag (Fig. 3). The MNS signatures were constructed very fast using a more efficient implementation described in [16]. The multimodal signatures for the database images were computed in 0.1 seconds on average and the query signature was computed in less than 0.1 seconds on a SUN Ultra Enterprise 450 with quad 400MHz UltraSPARC-II CPUs. The average signature size for the database images was 900 bytes.

In order to evaluate performance 13 "target" images containing the Irish colours were included in the database. The target images were manually selected from the same sequence as the database images and represented scenes of very different content. Objects containing the sought colours in the target images were often Irish flags sometimes occluded, non-rigidly deformed and/or of various sizes (Fig. 4). Sometimes, the frames were taken at shot transitions where video editing effects were apparent. Finally, illumination conditions changed dramatically between some of the frames resulting in completely different recorded colours. For example compare

Fig. 5. Sample daatabase images

image 3 with image 4(c) taken in the evening under very different light. Images 4 and 3 can be viewed in colour in [13].

The parameters involved in the computation of the signature and matching were not especially tuned for the task. Two consecutive randomised grid searches were performed with the same neighbourhood size (8×8 pixels) and the resulting multimodal neighbourhoods were merged into a larger set before clustering and computing the signature. Informative higher order features that are available at multimodal neighbourhoods with more than 2 modes were not exploited in the reported experiments. For the mean shift algorithm, a fixed kernel width of 25 units was used for the detection in the RGB space and 20 for the joint 6D space. Modes with support less than 10% of the neighbourhood were considered insignificant and therefore ignored. Low intensity modes (less than 5 percent of the luminance scale) were also not taken into account to improve stability especially in the case of relative colour feature matching. Although ratios from pixels with saturated (clipped) colours are not expected to be stable, we did not remove saturated colours for the reported experiments. The matching threshold was also fixed and was dependent only on the nature of the features used. For example, for RGB^2 feature matching, matching threshold was fixed to 100 for the proposed distance function (5).

4.2 Object Recognition Experiment

To compare MNS performance with results reported in the literature, we performed a well known colour object recognition experiment using a dataset collected by M. Swain. The database is publicly available [1] and has been used in a number of colour recognition experiments (e.g. [24, 9, 21]). The model image set consisted of 66 household objects imaged on black background under the same light (for a full colour image of the database see [24]) . The test set consisted of 32 images, a subset of model objects rotated, displaced or deformed (e.g. clothes). The test database and the corresponding model objects are shown in Fig. 6.

MNS performance evaluation was identical to Funt and Finlayson's [9] where ratio histogram matching was used for recognition. However, for that experiment, 11 model and 8 test images were removed from the database due to saturated pixels whereas we used all images. The same experiment was repeated by Park et al. [21] using a colour adjacency graph representation of image colour structure.

Computation of each MNS signature took 0.1 seconds on average. Image size was 128×90 pixels for both the model and test image sets. The average signature size was 150 bytes [16]. No image preprocessing, subsampling or smoothing was applied before signature computation. All internal parameters (mean shift kernel width, neighbourhood size etc) were exactly the same as those used for the image retrieval experiment.

5 Results

5.1 Image Retrieval

We first report results on the retrieval task from a database of sport images described in section 4. Database images were matched to the query image (Fig. 3) and sorted by their similarity to the query. Performance was evaluated according to the percentage of relevant images that were retrieved in the top 20 ranks of the retrieved list. In general, retrieval was very fast. A single signature match score was computed in approximately 1.5 ms i.e. the retrieval proceeds at 600 matches/sec on average. The results are presented in Fig. 8 as plots of percentage of relevant images as a function of the the number of retrieved images.

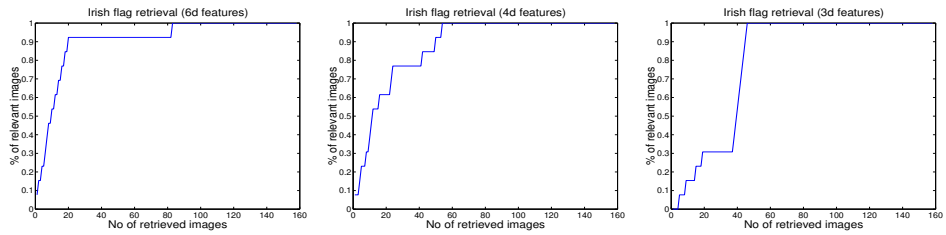
(a) Test objects used for the recognition experiment

(b) Model objects corresponding to the tests in (a)

Fig. 6. Sample test and model images from Swain's database

(a) (b) (c) (d)

Fig. 7. Examples of Swain's model images with very similar red-white regions: (a) clam chowder can, (b) chicken soup can, (c) mickey underwear and (d) red-white jumper



(a) (b) (c)

Fig. 8. Retrieval results for different colour invariant features using : (a) 6D RGB^2 features (b) 4D chromaticity features (c) 3D ratio features

(a) (b) (c) (d)

Fig. 9. Neighbourhoods that matched with the Irish flag query : (a) stripes on jumper (b) deformed flag in crowd (c) small flag on cap (I) (d) small flag on cap (II)

Retrieval results varied depending on the feature representation used. Assuming constant illumination for all images, 6-dimensional feature matching was applied which resulted in 12 out of 13 Irish images being in the top 20 ranks, a hit rate of 92.3%. However, the remaining image was ranked 83. The same retrieval experiment was repeated using the 4-dimensional chromaticity vectors. Hit rate was 61.5% but the worst rank was 53. We repeated the retrieval experiment for 3-dimensional ratio feature matching using only simple ratios as described in section 3.3. The worst match was 46 but hit rate was only 30.7%.

When matching based on absolute colour values, the colour constancy problem is apparent. An image with the query object of similar size (Fig. 4(c)) had a very low similarity score due to the significant change in the colour of illumination. Matching illumination invariant features, like the chromaticities or the ratios mentioned above, improved performance assigning a high rank to the previously missed image. Although the hit rate was not so good as in the first case, all relevant images were retrieved within a smaller subset of the retrieved image set. Clearly, there is a trade-off between the hit rate as defined above and the invariance to the illumination change. In the case where illumination colour was indeed constant, the higher dimensionality features benefited from their higher discriminative power. However, in changing illumination conditions, only relative invariant features were able to correctly rank the images even though discrimination was not as good.

Table 1. Comparative colour object recognition results for Swain's database

Method	Rank				Average Match	Number of test/model
	1	2	3	>3	Percentile	images
MNS	27	2	2	1	0.995	32 / 66
CC Colour Indexing	22	2	0	0	0.998	24 / 55
Colour Indexing	29	3	0	0	0.999	32 / 66
Hybrid graph	32	0	0	0	1.000	32 / 66

5.2 Colour Object Recognition

Assuming that illumination was kept approximately constant for all images in Swain's database the multimodal neighbourhood signature was tested using 6D RGB^2 feature matching. For each test object, signature dissimilarity from 66 model signatures (of the models described in section 4) was computed and the rank of the correct pair stored. For a single test object the recognition process took 0.28 seconds, i.e. 4 msec per image match. Speed was still real-time although slower than retrieval since the models were more complex than the Irish flag query image. To allow comparison with previous experiments, recognition performance of the algorithm was assessed in terms of the average match percentile. The match percentile for each image matched is defined as $\frac{N-r}{N-1}$ where N is the number of model images and r is the rank of the model image containing the test object.

Results are presented in Table 1. Recognition performance is compared with reported results for the colour indexing [24], colour constant (CC) colour indexing [9] and hybrid graph [21] representations respectively. Recognition using the MNS compared favourably to the other three algorithms with an average match percentile of 99.5% using the default MNS parameters.

The objects that were not classified as rank 1 include mainly objects with red-white colour boundaries (e.g. Fig. 7). Such object are common in Swain's database and their MNS signature is similar.

Histograms record areas (or relative areas if normalised) and have no problems discriminating between objects with almost identical colours but with different sizes of colour region. For Swain's database this property is beneficial, since most objects undergo only rotations and translations and have approximately the same scale. Consequently, MNS is outperformed, although the difference seems insignificant. In the presence of occlusion, object deformation or general view point change (e.g. as in the image retrieval experiment above) reliance on non-invariant and/or global property like area or relative area will negatively affect performance. The best reported result for this dataset was achieved by the hybrid colour adjacency graph which incorporates information about the spatial arrangement of colours in the image. Although the MNS representation allows for localisation of matching regions, we did not demonstrate this feature in the reported experiments.

6 Conclusions

In this paper, a novel approach to colour-based object recognition and image retrieval, the Multimodal Neighbourhood Signature (MNS) , was presented. The proposed method directly formulates the problem of representing object colour appearance by computing signatures of colour features derived from robust estimates of the modes of a local colour density function. From a multimodal neighbourhood signature, a number of invariants were computed to address changes in the imaging conditions within the application environment. In addition, by computing features

from image neighbourhoods, the MNS method facilitates region-based query specification and image retrieval.

We demonstrated our algorithm's performance on a region-based image retrieval task and a good (92%) hit rate was achieved in real time (600 image matches/sec on a SUN Ultra Enterprise 450 with quad 400MHz UltraSPARC-II CPUs). Relevant images were successfully retrieved regardless of background clutter, partial occlusion or non-rigid object deformation. In particular, very small regions were successfully matched like the small Irish flags on the swimmer's caps (Fig. 9). In addition, the trade-off between hit rate and illumination invariance was apparent in the reported experiments. Regarding colour object recognition, the MNS representation was tested on a standard dataset and compared favourably with three well known recognition algorithms. Very good performance (average match percentile 99.5%) was achieved with default settings, identical to those used in the image retrieval experiment. In general, the MNS signatures were concise and thus significant data reduction was achieved. An image was typically represented by a few hundred bytes, a few thousands for very complex scenes.

Future improvements to the algorithm include introducing a training/learning stage to efficiently exploit discriminative colour characteristics inherent to the database at hand, and a multiscale approach to compensate for scale changes. Selection of an appropriate distance for colour invariants, especially those taking the form of a ratio, should be investigated. Finally, we intend to study the potential of multimodal neighbourhoods with more than two modes for recognition and retrieval.

7 Acknowledgements

The ball image of Fig. 1 was from the image database of Simon Fraser University Canada available on-line [2]. The first author was supported by the Czech Ministry of Education under the grant VS96049. The second author acknowledges support from the Digital VCE, the EPSRC and the I. Latsis Foundation.

References

1. http://cs-www.uchicago.edu/users/swain/color-indexing/.
2. http://www.cs.sfu.ca/~colour/image_db/.
3. Silverman B.W. *Density Estimation for Statistics and Data Analysis.* New York:Chapman and Hall, 1986.
4. D. Comaniciu and P. Meer. Mean Shift Analysis and Applications . In *Proceedings of the International Conf. On Computer Vision*, pages 1197–1203, 1999.
5. M. Das, E. Riseman, and B. Draper. FOCUS: Searching for Multi-coloured Objects in a Diverse Image Database. In *IEEE Proceedings in Computer Vision and Pattern Recognition*, pages 756–761, 1997.
6. G. Finlayson, M. Drew, and B. Funt. Diagonal transforms suffice for color constancy. In *Proceedings of the International Conference on Computer Vision, Berlin*, pages 164–171, 1993.
7. K. Fukunaga and L. Hostetler. The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. In *IEEE Transactions in Information Theory*, pages 32–40, 1975.
8. B. Funt, K. Barnard, and L. Martin. Is Machine Colour Constancy Good Enough? In *Proceedings of the 5th European Conference on Computer Vision*, pages 445–459, 1998.
9. B. Funt and G. Finlayson. Color Constant Color Indexing . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):522–529, 1995.
10. T. Gevers and W. M. Smeulders. Color-based Object Recognition. *Pattern Recognition*, 32(3):453–464, 1999.

11. G. Healey and D. Slater. Global Color Constancy - Recognition Of Objects By Use Of Illumination Invariant Properties Of Color Distributions . *Journal Of The Optical Society Of America A-optics Image Science And Vision*, 11(11):3003–3010, 1994.
12. D. Jacobs, P. Belhumeur, and R. Basri. Comparing Images Under Variable Illumination. In *IEEE Proceedings in Computer Vision and Pattern Recognition*, pages 610–616, 1998.
13. D. Koubaroulis, J. Matas, and J. Kittler. MNS: A Novel Method for Colour-Based Object Recognition and Image Retrieval. Technical Report VSSP-TR-6/99, University of Surrey, December 1999. (available at http://www.ee.surrey.ac.uk/Personal/D.Koubaroulis/thesis/pub/koubaroulis-tr699.ps.gz).
14. E. Land and J. McCann. Lightness and the retinex theory. *Journal of the Optical Society of America*, 61(1):1–11, 1971.
15. J. Matas. *Colour Object Recognition*. PhD thesis, University Of Surrey, 1995.
16. J. Matas, D. Koubaroulis, and J. Kittler. Performance Evaluation of the Multi-modal Neighbourhood Signature Method for Colour Object Recognition. In *Proceedings of the Czech Pattern Recognition Workshop, Perslac, Czech Republic*, pages 27–34, 2000. (available at http://www.ee.surrey.ac.uk/Personal/D.Koubaroulis/thesis/pub/matas-cprw00.ps.gz).
17. K. Messer, J. Kittler, and M. Kraaijveld. Selecting Features for Neural Networks to Aid an Iconic Search Through an Image Database . In *Proceedings of the IEE 6th International Conference on Image Processing and Its Applications*, pages 428–432, 1997.
18. F. Mindru, T. Moons, and L. Van Gool. Recognizing Color Patterns Irrespective of Viewpoint and Illumination. In *Proceedings of the Computer Vision and Pattern Recognition,Fort Collins, Colorado*, pages 368–373, 1999.
19. K. Nagao and W. Grimson. Recognizing 3D Objects Using Photometric Invariant. Technical report, Massachusetts Institute of Technology Artificial Intelligence Lab, 1995.
20. S. Nayar and R. Bolle. Reflectance Based Object Recognition. *International Journal of Computer Vision*, 17(3):219–240, 1996.
21. K. Park, Il-Dong Yun, and Sang Uk Lee. Color Image Retrieval Using a Hybrid Graph Representation. *Journal of Image and Vision Computing*, 17(7):465–474, 1999.
22. Y. Rui, T.S. Huang, and S.-F. Chang. Image retrieval: Current techniques, promising directions, and open issues . *Journal Of Visual Communication And Image Representation*, 10(1):39–62, 1999.
23. R. J. Smith and S.-F. Chang. Integrated Spatial and Feature Image Query . *Multimedia Systems*, 7(2):129–140, 1999.
24. M. J. Swain and D. H. Ballard. Color Indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.