COMPACTED DITHER PATTERN CODES OVER MPEG-7 DOMINANT COLOUR DESCRIPTOR IN VIDEO VISUAL DEPICTION

Lochandaka Ranathunga, Roziati Zainuddin, Nor Aniza Abdullah
Faculty of Computer Science and Information Technology,
University of Malaya, 50603, Kuala Lumpur, Malaysia
lochandaka@perdana.um.edu.my, roziati@um.edu.my, noraniza@um.edu.my

ABSTRACT

Reduction of feature space of visual descriptors has become important due to the 'curse of dimensionality' problem. This paper reports the efficiency and effectiveness of the Compacted Dither Pattern Code (CDPC) combined with the Bhattacharyya classifier over MPEG-7 Dominant Colour Descriptor (DCD). Both the CDPC and DCD syntactic features use a compact feature space for colour representation. The algorithmic comparison between the two is presented in this paper, and demonstrates that there are several competitive advantages of CDPC in feature extraction and classification stages when compared to MPEG-7 DCD. The embedded texel properties, spatial colour arrangements, high compactness, and robust feature representation of CDPC have proven its effectiveness in our experimental study. Visual description experiments were conducted for ten irregular shapes-based visual concepts in videos with three setups namely CDPC with Bhattacharyya classifier, DCD without spatial coherency and DCD with spatial coherency. The visual descriptions were performed with the TRECVID 2007 development key frame dataset. The experimental results are presented in terms of three common performance measures. The results show that CDPC with Bhattacharyya classifier provides a good generalised performance for irregular shapes-based visual description as compared to the other experimental setups.

Keywords: Bhattacharyya classifier, Compacted Dither Pattern Code, Dominant Colour Descriptor, Generalised Lloyds Algorithm, MPEG-7, Video Content Analysis, Visual Concept Depiction, Visual Features.

1.0 INTRODUCTION

Due to increasing bulk of data in video repositories, automatic extraction of visual information from videos is widely required. The vivid successes of multimedia technologies has resulted in rapid increases in image and video archives including media publications, internet publications and organizational applications. There is a considerable effort of academic research in multimedia information retrieval, but relatively little impact caused to the industry with some exceptions such as video segmentation [1].

The popular video search engines such as Google, Blinkx, and YouTube provide access to their repositories based on textual annotations. The indices of these search engines are based on basic metadata [2]. The incorporation of visual information in enriching video description is one challenge focused on by the research community using different approaches. In addition, the visual analysis yields more robustness [2]. Therefore, there is a rising need for the development and improvement of algorithms and techniques to browse, search, retrieve and manage the visual content [3,4]. An automated video visual content description can be viewed as a three stages of processes. In order, these stages are visual feature extraction, feature classification, and visual concept depiction. The effectiveness of each stage depends on the success of the previous stage.

The negative impact of high dimensional feature space demands that low dimensional robust visual features for visual content depiction be found [5,6,7]. In reducing feature space, it is a difficult problem not to sacrifice visual depiction quality [8]. The approach devise in this study uses low dimensional feature space while retaining the useful information. Our approach has given generalised improved results in the presence of diverse visual qualities and unbalanced concept distributions of the dataset.

Our methodology operates on video visual regions and extracts micro-level compact patterns. This enables us to study the behaviour and relationship between classification accuracy and computational cost compared to the well-known compact visual descriptor namely MPEG-7 Dominant Colour Descriptor (DCD). While reducing the feature space, our descriptor is able to maintain an improved performance compared to DCD.

2.0 RELATED STUDIES

Even though the research community in the field of multimedia information retrieval has generally admited that there are no general solutions, there are special cases when the general problem can be reduced to a smaller niche problem where high accuracy and precision can be quantitatively demonstrated [1]. This is true for visual information depiction. Video visual information retrieval accuracy and performances depend on syntactic features and the classifier being used. By identifying this fact, MPEG has devised a special category under visual colour features in their MPEG-7 Multimedia Content Description Interface scope [9]. Researchers have experimented on feature extraction of video with colour histograms, textures, motion information and shapes [3, 10]. However in visual feature extraction the colour is the foremost feature used in deriving high-level visual concepts [11]. Therefore intended visual feature characterization should generalize in terms of illumination variance, colour variance and geometrical variances [3, 10, 12]. Researchers have used colour and geometrical distributions to improve results [13, 14, 15]. Studies in [15] and [16] have used a mixture of colours and texture properties disregarding their efficiency in classification.

The recent trend is towards interest in region-based feature extractors and descriptors [2]. Salient point of extraction algorithms have generally shown improved results in terms of retrieval accuracy and storage space of feature vectors as compared to the global feature approaches [17, 18]. In general the local feature methods are efficient for Object-based retrieval [12]. Study in the area of Local Interest Points (LIP) has been explored and the defects of regional and interest point based methods are mentioned in [8, 17, 18, 19]. Local feature-based descriptors such as SIFT (Scale Invariance Feature Transform) and Gradient Location Orientation Histogram (GLOH) contain high dimensionality in feature space which causes major negative impact in computation [8, 18, 19]. The local and regional features are used in combination with other features for better performances in some studies [16]. This has proliferated the comparison overheads and calculations.

MPEG-7 focuses on content description of various multimedia asserts including images, videos, audios, speeches, graphics, and their combinations [9]. The characteristics of the colour arrangement in the visual description relates to the perception, coherency and spatial distribution [20]. In visual content description and retrieval, MPEG-7 visual standard for content description has provided a range of descriptors corresponding to common features including colour, texture, shape, and motion. Varieties of MPEG-7 colour descriptors are popular in visual concept detection and description [20, 21]. The Dominant Colour Descriptor (DCD) is one of the main descriptor which has low dimensionality and computational cost [11, 20, 21]. The DCD has been experimented on in many visual depiction research projects [22, 23].

2.1 Dominant Colour Descriptor (DCD)

The key target of the DCD development is to achieve fast and efficient visual depiction and retrieval [21]. The DCD uses local spatial colour distribution at global level. The purpose of this descriptor is to provide effective, compact and intuitive description of the representative colours of an image or image region [11, 21]. This descriptor uses the salient colour in visual regions in pixel domain using a colour clustering process.

Generally it uses the CIE-LUV colour space in colour clustering. The DCD clusters a given visual region into a small number of representative colours. The feature descriptor consists of the representative colours, their percentages in the region, spatial coherency of the dominant colours, and colour variances for each dominant colour [11,21]. Usually the number of reprehensive colours is less than or equal to 8. This size of feature vector has reduced the dimensionality of the feature space. The visual classification is similar to the quadratic colour histogram distance measure which defined in this descriptor. This descriptor avoids the high-dimensional indexing problems associated with the traditional colour histogram [11]. The feature vector consists of the colour index (c_i), percentage (p_i), colour variance (v_i) and special coherency (s) where the last two are optionals [24]. It is defined by,

$$DCD = \{(c_i, p_i, v_i), s\}\ , i = 1, .., N \qquad (1)$$

where N is the number of colours and $\sum_{i=1}^{N} p_i = 1$

The spatial coherency is a single number that represents the overall spatial homogeneity of the dominant colours in the visual region.

2.2 Dominant Colour Extraction

The extraction of dominant colours uses the generalised Lloyd algorithm [11, 25]. This algorithm is a popular method for K-mean clustering. The algorithm performs clustering by minimising the distortion *D_i* in each cluster *i*. And the steps of the algorithm iterate until there is no movement found in points among clusters.

$$D_i = \sum_n v(n) \| x(n) - c_i \|^2 \qquad x(n) \in C_i \tag{2}$$

where c_i is the centroid of cluster C_i, *x(n)* is colour vector at a pixel, and *v(n)* is perceptual weight for pixel *n*. The perceptual weights are calculated from the local pixel statistics to account for the fact that human vision perception is more sensitive to changes in smooth regions than in textured regions. The effective use of perceptual weight was experimented in [26].

There are some studies conducted with the following modification in extraction of dominant colours.

$$c_i = \frac{\sum_n v(n).x(n)}{\sum_n v(n)} \qquad x(n) \in C_i \tag{3}$$

A simple connected component analysis is performed to identify groups of pixels of the same dominant colour that are spatially connected [11]. The normalized average number of connecting pixels of the corresponding dominant colour is computed. A masking window of *3 x 3* is utilized to measure the spatial coherence of a given dominant colour. The overall spatial variance is a linear combination of the individual spatial variances with the corresponding p_i percentages being the weights [11]. The spatial variance is quantized to 5 bits, where 31 means highest confidence and 1 means no confidence. 0 is used for cases where it is not computed [11]. The impotency of spatial homogeneity in visual matching is emphasized in a number of research studies [27, 28].

2.3 DCD Matching Process

MPEG-7 has given directions to compute the similarity of two dominant colour descriptors. There is one overall spatial coherency value for the given image region and several groups of (c_i, p_i, v_i) for the corresponding dominant colours. It can be used to compute the visual difference between images based on colour. It is a quadratic type of distance measure similar to histogram distance [11, 29].

Consider two DCD feature vectors,

$F1 = \{(c_{1i}, p_{1i}, v_{1i}), s_1\}, (i = 1,2 \ldots N_1)$ and *$F2 = \{(c_{2i}, p_{2i}, v_{2i}), s_2\}, (i = 1,2 \ldots N_2)$*

Considering optional variance and spatial coherence, the dissimilarity of two feature vectors *D(F1,F2)* is described as,

$$D^2(F1,F2) = \sum_{i=1}^{N_1} p_{1i}{}^2 + \sum_{j=1}^{N_2} p_{2j}{}^2 - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} 2a_{1i,2j} p_{1i} p_{2j} \tag{4}$$

The $a_{k,l}$ represent the similarity between two colours of c_k and c_l.

$$a_{k,l} = \begin{cases} 1 - d_{k,l} / d_{\max} & d_{k,l} \le T_d \\ 0 & d_{k,l} > T_d \end{cases} \tag{5}$$

where,

$$d_{k,l} = \left\| c_k^{\ 2} - c_l^{\ 2} \right\| \tag{6}$$

Is the Euclidean distance between two colours and T_d is maximum for two colours considered similar.

The d_{max} in (5) maintain the following equality to produce consistency within colour separation of two dominant colours.

$$d_{\max} = \alpha T_d \tag{7}$$

A normal value specified for T_d is between 10 – 20 in the CIE-LUV colour space and α is considered to be between 1.0 – 1.5.

The above dissimilarity can be further modified with a linear combination of the spatial coherency. Then the dissimilarity including spatial coherency D_s is defined by

$$Ds = w_1(s_1\text{-}s_2)\ .\ D(F1,F2) + w_2\ .\ D(F1,F2) \tag{8}$$

Where, s_1 and s_2 are spatial coherency values for two descriptors. According to MPEG-7 w_1 and w_2 are fixed weights, with recommended settings to 0.3 and 0.7 respectively [25].

In both the situations of equation (4) and (8), when the distance is lower, visuals are considered to be more similar. However this dissimilarity measure has to be performed in each matching situation of a query feature vector against trained feature vectors. So in both situation of finding the similar and dissimilar visual regions the calculation cost is consistent. As most of the specification given for the CIE-LUV colour space, there is a need of colour space conversion at the beginning. This is another task included in dominant colour feature extraction. The optional spatial coherency value indicates the level of homogeneity of distribution of all extracted dominant colour within a given region. But it is not a value to determine micro-level colour pattern arrangements or texture properties. In the case of using colour variance optional field, the dissimilarity measure utilises a mixture of Gaussian distributions as a colour distribution and this makes dissimilarity calculations more complex, with higher overhead [11, 25].

In this study we devise a novel compact form of representation of the colours of video visuals as a syntactic feature. This compact form reduces the visual depth and density by keeping feature space in low dimensionality which can be considered as the probability distribution of a pattern set in a given region. This form of visual feature provides definite advantages in visual classification.

3.0 COMPACTED DITHER PATTERN CODES

Most graphical applications utilise colour dithering to reduce chromatic depth and spatial density. The visual semantics are not affected by those processors. Here we formulate a modified low dimensional visual pattern code set with reduction of chromatic depth and spatial density for irregular shape based video visual detection and description. It is called “Compacted Dither Pattern Code” (CDPC) and can be used as a syntactic visual feature to describe semantic visual concepts with irregular shapes. In this mechanism of video visual content analysis, probability distribution of CDPCs is used to identify and classify visual concepts. CDPC has helped to reduce the complexity and uncertainty accompanied with high colour depth and spatial resolution of bulky videos. This syntactic feature is accompanied with neighbourhood (spatial) chromatic information in micro pattern levels. It contains some properties of texels as it represents combinational chromatic information over a tiny square space.

The objectives of selecting CDPC as a syntactic feature in video content extraction are to address video colour pattern clustering using combinatorial pattern set, bring down higher level colour depth and spatial density in videos to addressable and distinguishable pattern levels, reduce the computational cost and complexity of clustering and depiction phases by using one knowledge scheme, and incorporate neighbourhood information and texel properties into the same syntactic feature extractor.

When detecting video scene concepts with irregular shapes-based visual concepts need the assistance of colour based syntactic feature extractors. However in order to detect those sceneries, shape, edge and salient features are ineffective. In video frames, pixel colours are different in the same region even though we see them as one colour. This can be taken as distribution of colour grid patterns or dither patterns when converting these grid patterns in an orderly manner.

3.1 CDPC Feature Extraction

The *r,g,b* additive colours were presented in *z* number of colour levels. The CDPC grid size of *l=2x2* square dither colour points, in which each of dither point in CDPC grid represents, t^2 pixels of block. So the number of different colour dither pattern grids created will be given by,

$$N = {}^{z}C_4 + {}^{z}C_2 + \sum_{r=0}^{2} {}^{z}C_r.(z-r) \tag{9}$$

Where $z >= 4$. By considering smaller t^2 blocks and spatial blending effects, the different internal colour arrangements within CDPC grid can be excluded. The above equation represents the selection of the number of elements in CDPC set.

3.2 Steps of CDPC feature extraction

Step1: The video frames were treated with t^2 mask of colour averaging filter as in equation (10). A set of coordinates of selected grid of pixels *S, (x,y) Є S* is point pixel represents the whole block as a part of dither pattern, and $N_G = t^2$ is the number of pixels in the block.

$$f_{R,G,B}(x,y) = \frac{1}{N_G} \sum_{p,q \in S} f_{R,G,B}(p,q) \tag{10}$$

The $f_{R,G,B}(x,y)$, 3x1 column vector contains the representation of RGB chromatic data of the block.

Step 2: Use a three dimensional vector quantization to produce an individual dither point for each t^2 pixels block. If *Z'* is the colour component fragmentation factor, then it satisfies the *Z' * (Integer value) = 255*, equality.

$$f_{R,G,B}(x,y) = \lfloor ((f_{R,G,B}(x,y)/Z') + 0.5) \rfloor * Z' \tag{11}$$

By using (11), all the blocks are transformed into dither levels, where $\lfloor\ \rfloor$ function will return the integer part of the value. The above quantization prepares the video frame colours into *z* colour levels.

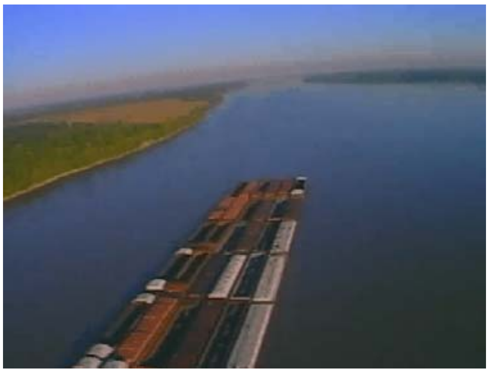


Fig. 1a: Video frame extracted from a sequence

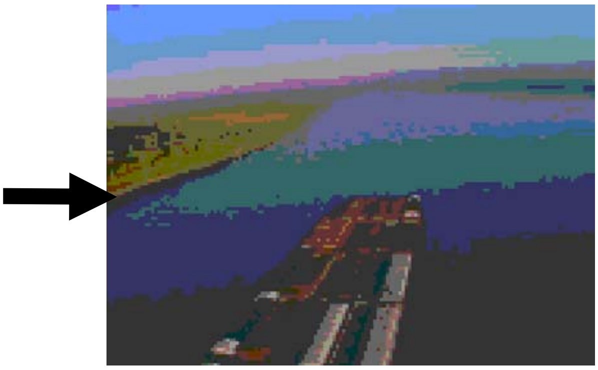


Fig. 1b: Visualization of extracted video frame after performing step 1 and step 2

Step 3: Extraction of four point codes at *(x,y), (x+t,y), (x, y+t)* and *(x+t, y+t)* for each *x,y* with *2t* intervals, where *x+2t<=given spatial width & y+2t<=given spatial height* in each case.

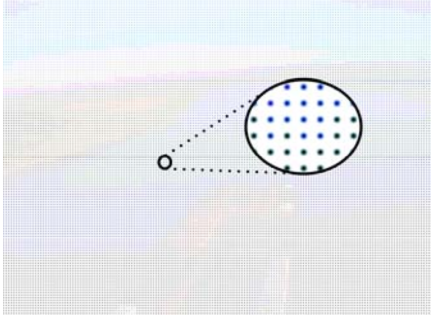
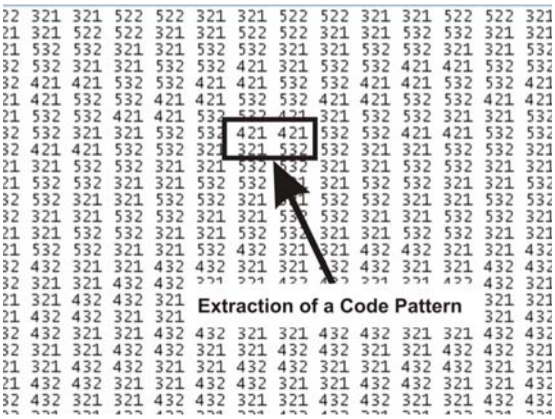


Fig. 2: Visualization of extracted video frame after performing points extraction in step 3. Magnified circle area shows reduction of spatial density

Step 4: Obtain Dither Code (*DC*) components for each dither grid representation point by following equation (12).

$$DC_{R,G,B}(x,y) = \lfloor ((f_{R,G,B}(x,y)/Z') + 0.5) \rfloor \tag{12}$$

Here $DC_{R,G,B}(x,y)$ represents a 3x1 column vector with each element in the range of *0-255/Z'*. The resultant channel codes can be concatenated or transposed $[DC_{R,G,B}(x,y)]^T$, and it generates an individual dither pattern point code which contains three values according to the dithered chromatic levels. Fig. 3 illustrates the resultant code patterns and reduction of chromatic depth after performing step 4.



```
22 321 321 522 522 321 321 522 522 321 321 522 522 321
21 321 522 522 321 321 522 522 321 321 532 532 321 321
21 532 532 321 321 532 532 321 321 532 532 321 321 532
32 532 321 321 532 532 421 321 532 532 421 421 532 532
32 421 421 532 532 421 421 532 532 421 421 532 532 421
21 421 532 532 421 421 532 532 421 421 532 532 421 421
21 532 532 421 421 532 532 421 321 532 532 321 321 532
32 532 321 321 532 532 421 421 532 532 421 421 532 532
32 421 421 532 532 321 321 532 532 321 321 532 532 321
21 321 532 532 321 321 532 532 321 321 532 532 321 321
21 532 532 321 321 532 532 321 321 532 532 321 321 532
32 532 321 321 532 532 321 321 532 532 321 321 532 532
32 321 321 532 532 321 321 532 532 321 321 532 532 321
21 321 532 532 321 321 532 532 321 321 532 532 321 321
21 532 532 321 321 532 432 321 321 432 432 321 321 432
32 432 321 321 432 432 321 321 432 432 321 321 432 432
32 321 321 432 432 321 321 432 432 321 321 432 432 321
21 321 432 432 321                                321 321
21 432 432 321 321                                321 432
32 432 321 321 432 432 321 321 432 432 321 321 432 432
32 321 321 432 432 321 321 432 432 321 321 432 432 321
21 321 432 432 321 321 432 432 321 321 432 432 321 321
21 432 432 321 321 432 432 321 321 432 432 321 321 432
32 432 321 321 432 432 321 321 432 432 321 321 432 432
```

Fig. 3: Visualization of code patterns generated from extracted video frame region. The *Z'* used here is 51.0 and each chromatic channel varies from 0-5.

Step 5: Generate CDPC by arranging each extracted four codes set in descending order. This is to exclude internal permutations within the pattern. When considering the tiny size of the grid and the spatial blending effect each internal permutation is considered to be represented by an element of a selected CDPC set.

```
<p1>532</p1><p2>421</p2><p3>421<p3><p4>321</p4>
<Prob>0.2104</prob>
```

Fig. 4: Demonstration of a Prepared of CDPC with the extracted code pattern in Fig. 3

Fig. 4 illustrates the preparation of CDPC according to the extracted pattern code in Fig. 3. Further the Fig. 4 shows a CDPC element structure according to an XML schema after performing step 5. The value for the element <prob> (probability) is calculated in step 7.

Step 6: Calculate the population of each CDPC category in a given spatial area. As each pattern has a strict sequence, the sorting and calculation can be done fast.

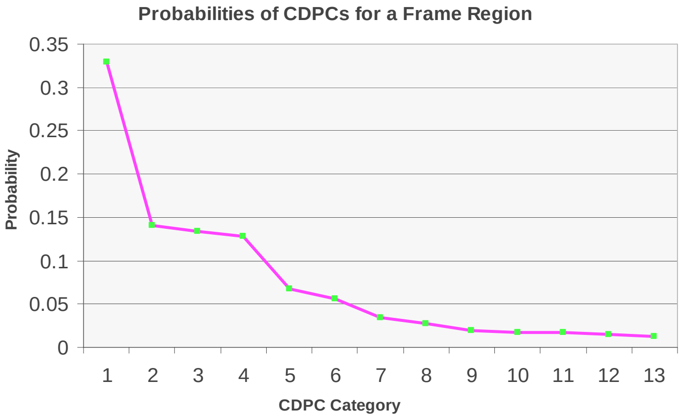


Fig. 5: Visualisation of probability distribution of a region of extracted frame

Step 7: Obtain probability of each CDPC category in the specified spatial region. The resultant probability distribution is limited to maximum of 20 patterns with highest probabilities. The CDPCs are arranged in the highest to lowest probability. One example for CDPC probability distribution is given in Fig. 5.

4.0 VIDEO SCENE CONCEPTS DETECTION AND CLASSIFICATION

Fig. 6 illustrates the components of Video Scene Concepts Detection and Classification system which uses CDPC as a syntactic feature. This system comprises of two sub systems namely, Training System and Scene Concept Depiction System. The Training System is used for constructing Knowledge Repository for different scene concepts. Here, the generated Video Concepts' Knowledge Repository which contains probability distributions of CDPC of different scene concepts.

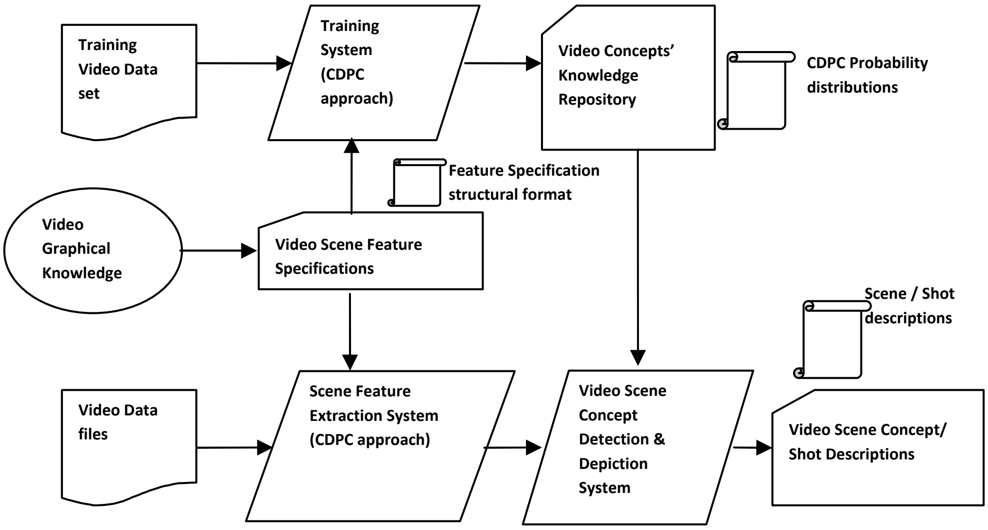

Training Video Data set
Training System (CDPC approach)
Video Concepts' Knowledge Repository
CDPC Probability distributions
Feature Specification structural format
Video Graphical Knowledge
Video Scene Feature Specifications
Scene / Shot descriptions
Video Data files
Scene Feature Extraction System (CDPC approach)
Video Scene Concept Detection & Depiction System
Video Scene Concept/ Shot Descriptions

Fig. 6: Components of Video Scene Concepts Detection and Classification System

The Video Concepts' Knowledge Repository was constructed according to a well structured XML knowledge schema and therefore contains XML data files. When generating Video Concepts' Knowledge Repository, the directions for indicating video concepts were given by using Video Scene Feature Specifications which is an XML schema based structural format with spatial attributes, temporal attributes, file properties and CDPC attributes.

The Feature Specification structural format gives the required attributes for video scene processing in both Training and Depiction. It uses video frame of 16 rectangular sections with processing of selected section/s. The Scene Concept Depiction System has two stages of a processing system. They are Scene Feature Extraction System and Video Scene Concept Detection and Depiction System. The Scene Feature Extraction System is used to extract the probability distributions of CDPC according to the given specification by the Feature Specification structural format. Then the extracted probability distributions of CDPC for each video frame/section will be used by Video Scene Concept Detection and Depiction System, to classify the extracted CDPC probability distribution with the help of Knowledge Repository. In classifying video scene concepts the Bhattacharyya coefficient was used [30], because CDPC probability distributions are multinomial probability distributions and Bhattacharyya classifier is converged with classification decision into one numerical value. Bhattacharyya coefficient value helps in the identification of most accurate match of the labelled pattern distribution. Finally the Video Scene Concept/Shot Description component generates visual concept descriptions of video frames in accordance with Scene/Shot description schema. The Scene/Shot description schema was developed according to MPEG-7 video visual description guidelines.

5.0 VIDEO SCENE CONCEPT CLASSIFICATION USING CDPC

If a possible set of CDPC has *N* number of codes and the CDPCs set is represented by *E*, then all elements of *E* can be represented as,

$$E=\{CDPC(1), CDPC(2), .. CDPC(N)\} \tag{13}$$

where *CDPC(j)* is any CDPC in *E*.

Let trained knowledge repository has Λ set of video scene concepts with different CDPC probability distributions,

$$\Lambda = \{\lambda_1, \lambda_2 , \lambda_3 ,\ldots.. \lambda_n \} \tag{14}$$

where λ_i is any trained video scene concept in a Knowledge Repository with CDPC probability distributions.

A corresponding set of CDPC probabilities for the λ_i is given by $\Phi(j) \epsilon \Phi$ and elements are defined as,

$$\Phi(j) = \{\alpha_1 , \alpha_2 , \alpha_3 , \ldots.. \alpha_N\} \tag{15}$$

where α_j corresponds to the probability of *CDPC(j)* in trained knowledge concept.

Similarly, let us take a set of CDPC probability distribution extracted from any video frame section/s as Φ' and elements are defined as,

$$\Phi' = \{ \alpha'_1 , \alpha'_2 , \alpha'_3 , \ldots.. \alpha'_N\} \tag{16}$$

where α'_j corresponds to the probability of *CDPC(j)* in video frame section/s

In the classification of unknown video scene concept into known video scene concept, it is required to calculate the best match probability distribution among the Video Concepts' Knowledge Repositories.

$\Phi(j)$, Φ' can be considered as two multinomial populations characterized by two sets of probability populations, $\{ \alpha_1 , \alpha_2 , \alpha_3 , \ldots.. \alpha_N\}$ *and* $\{ \alpha'_1 , \alpha'_2 , \alpha'_3 , \ldots.. \alpha'_N\}$.

As $\sum \alpha_j = 1$ and $\sum \alpha'_j = 1$ by using Bhattacharyya divergence between two multinomial population [30], then

$\{\sqrt{\alpha_1}, \sqrt{\alpha_2}, ... \sqrt{\alpha_N}\}$ and $\{\sqrt{\alpha'_1}, \sqrt{\alpha'_2}, ... \sqrt{\alpha'_N}\}$ can be considered to be the direction cosine of two straight lines through the origin in a *N*-dimensional space. The square of the angle between these two lines can be considered to be an appropriate measure of the divergence between two multinomial populations.

Thus, if the measure of divergence can be denoted by Δ^2, then

$$Cos\ \Delta = \sqrt{\alpha_1 \alpha'_1} + \sqrt{\alpha_2 \alpha'_2} + \sqrt{\alpha_3 \alpha'_3} + ... + \sqrt{\alpha_N \alpha'_N} \tag{17}$$

This can be written as

$$4 Sin^2 \Delta/2 = (\sqrt{\alpha_1} - \sqrt{\alpha'_1})^2 + (\sqrt{\alpha_2} - \sqrt{\alpha'_2})^2 + (\sqrt{\alpha_3} - \sqrt{\alpha'_3})^2 + ... + (\sqrt{\alpha_N} - \sqrt{\alpha'_N})^2 \tag{18}$$

From this later expression it is evident that when $\sqrt{\alpha_j} = \sqrt{\alpha'_j}$ (for *j=1,2,...N), Δ* vanishes. Also when *Δ* vanishes, it can be conversely taken that $\sqrt{\alpha_j} = \sqrt{\alpha'_j}$ *(j=1,2,...N),* that is, the two populations are identical.

The expression

$$Cos \Delta = \sum_{i=1}^{N} \sqrt{\alpha_i \alpha'_i} \tag{19}$$

is called Bhattacharyya coefficient in discrete probability distributions.

So the measure of divergence of two probability distributions can be represented as,

$$\Delta^2 = [Cos^{-1}(\sum_{i=1}^{N} \sqrt{\alpha_i \alpha'_i})]^2 \tag{20}$$

This measure was proposed by Bhattacharyya and is known in the statistics literature (as also in some application areas such as physics and computer science) as Bhattacharyya's distance [31].

The Bhattacharyya coefficient is 1 when two populations are identical and 0 when two populations are totally different. When the coefficient value is closer to 1, then two probability distributions are also more similar. Therefore the Bhattacharyya coefficient was used in this video scene concept classification.

Therefore, based on the Bhattacharyya coefficient, a threshold conditional value was set to identify positive CDPC probability distributions in Knowledge Repository as compared to unknown CDPC probability distribution. Based on positive CDPC probability distributions set of $\Omega \subset \Phi$, the unknown CDPC probability distribution is classified into the *Φ(i)* (ϵ Ω) which gain the highest Bhattacharyya coefficient. Therefore according to classified *Φ(i),* the Video Scene Concept λ_i, will be selected for depiction.

6.0 COMPUTATIONAL COMPARISON OF DCD WITH CDPC DESCRIPTOR

MPEG-7 DCD design goals have included compaction of feature representation and improvement of the efficiency of visual content depiction. During visual feature extraction stage, DCD usually transforms colour space into Luminance, Chrominance spaces (CIE-LUV is more often) whereas CDPC does not use any colour space transformation. The dominant colour clustering or selecting of dominant colours uses generalised Lloyd algorithm. The Lloyd's algorithm is an iterative algorithm which uses equation (2) [11]. It uses Euclidian distance measure in each point classification. The algorithm constructs a Veronoi diagram for each iteration process [32]. The steps of generalised Lloyd have integrated determination of centroid coordinates of each cluster, and determination of distance to each point from centroids and group points according to minimum distance. The algorithm iterates until the Veronoi diagram becomes stable. In *d* dimensional space with *k* clusters the complexity goes to $O(n^{dk})$, where *n* is the number of points [32]. In CDPC based approach, it utilises neighbourhood average of t^2 pixels blocks which are the extracted and quantised CDPC points according to steps 3 & 4 in section 3.2 and CDPC patterns are arranged according to their probabilities. CDPC feature extraction complexity can go to $O(n^2)$ in the worst case, which may be due to quick sorting algorithm. So CDPC feature extraction is not effected by dimensionality factors of chromatic data. Note that DCD uses floating point arithmetic in feature extraction while CDPC uses integer arithmetic.

DCD feature vector without optional spatial coherency and colour variance has four dimensions per element. The number of dominant colours can vary up to eight numbers. CDPC feature vector can be represented in two dimensional spaces as shown in Fig. 5. The number of elements can be up to twenty patterns per region. When comparing dimensions of the feature vectors, the CDPC gives a more compact form.

The process of concept classification with DCD uses equation (4) without optional fields. DCD classifier varies as in equation (8) when considering spatial coherency optional value. According to equation (4) to (7) in DCD similarity measure, all the feature points in the extracted features and targeted features are employed in calculation. CDPC with Bhattacharyya classifier selects only similar patterns for the calculation by comparing extracted CDPCs with targeted CDPCs. As CDPC uses strict sequence in pattern, the comparison is much easier. Bhattacharyya classifier uses only a simple calculation as in equation (19) to generate decisive coefficient. Both DCD distance measure and Bhattacharyya classifier uses floating point arithmetic in classification. When considering optional spatial coherency value, the calculation overhead of DCD classifier increases.

In summing up the comparison of DCD and CDPC based methods, it can be seen that the feature extraction complexity of DCD is expensive than CDPC feature extraction. Further, the lower dimensionality and strict structure of CDPC feature vector is an advantage in the classification. The classification process of DCD with distance measure is computationally expensive than CDPC with the Bhattacharyya classification process.

7.0 EXPERIMENTAL SETUP

To evaluate the performances of CDPC based video visual descriptor and DCD, we have used colour videos and movie key frames from TRECVID 2007 development data set [33]. A total of 2000 random shots were created using the above data set inline with the data preparation processes of data set [34]. There were 10 irregular shapes based visual concepts namely, sky, sea, mountains, vegetation, buildings, roads, water resource, fire and snow were selected for depiction and training. The selected concept set included is visually identical to concept couples such as sky and sea, sea and water resource, sky and snow. So this is an opportunity to evaluate the distinguishable power of the descriptors as well. In a preliminary experiment an optimal number of samples was identified to be used for training using CDPC syntactic feature. The required knowledge repository for the system was created by using 30-35 numbers of training video samples from each concept. The balance of training among concepts was maintained within the experiment. The development of knowledge base was done by using special XML knowledge schema structure. The outline structure of knowledge schema is shown in Fig.7. Therefore the video concepts knowledge repository in Fig. 6 has the same outline structure.

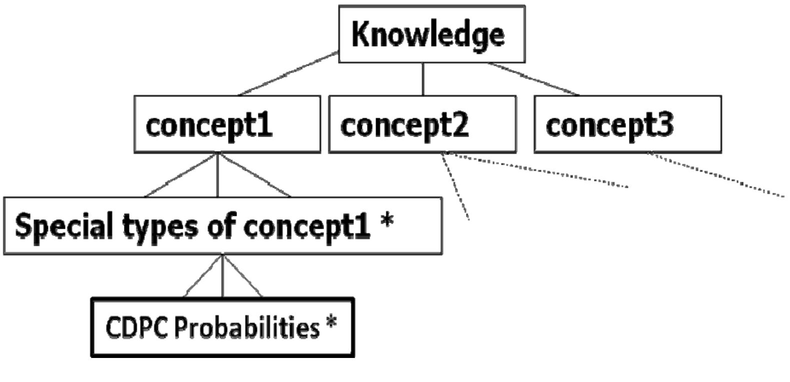
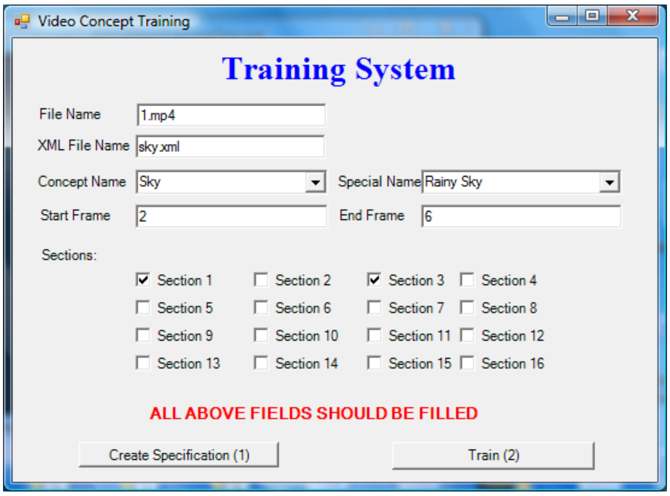


Fig. 7: Outline of knowledge schema structure
*denotes multiple items under hierarchy

In the above knowledge schema structure, each concept holds a number of special situations of the same visual concept. Each of those special situations is labelled against its own CDPC probability distributions. Therefore this schema supports not only main concept depiction, it also can provide extended description with concept's special events. A special interface was used to construct the trained knowledge repository according to the knowledge schema. The training system interface is shown in Fig. 8. The same training set and the training interface was used for the DCD feature storage generation. The DCD feature vector was trained with spatial coherency option for later use. The training system divides each video frame into 16 rectangular sections. This division mechanism was used in concept depiction as well in which the depiction is done for 16 different sections per video frame. This scheme has enabled different concepts appeared in different regions of video frames to be utilised separately in training.



Video Concept Training

Training System

File Name: 1.mp4

XML File Name: sky.xml

Concept Name: Sky

Special Name: Rainy Sky

Start Frame: 2

End Frame: 6

Sections:

☑ Section 1 □ Section 2 ☑ Section 3 □ Section 4

□ Section 5 □ Section 6 □ Section 7 □ Section 8

□ Section 9 □ Section 10 □ Section 11 □ Section 12

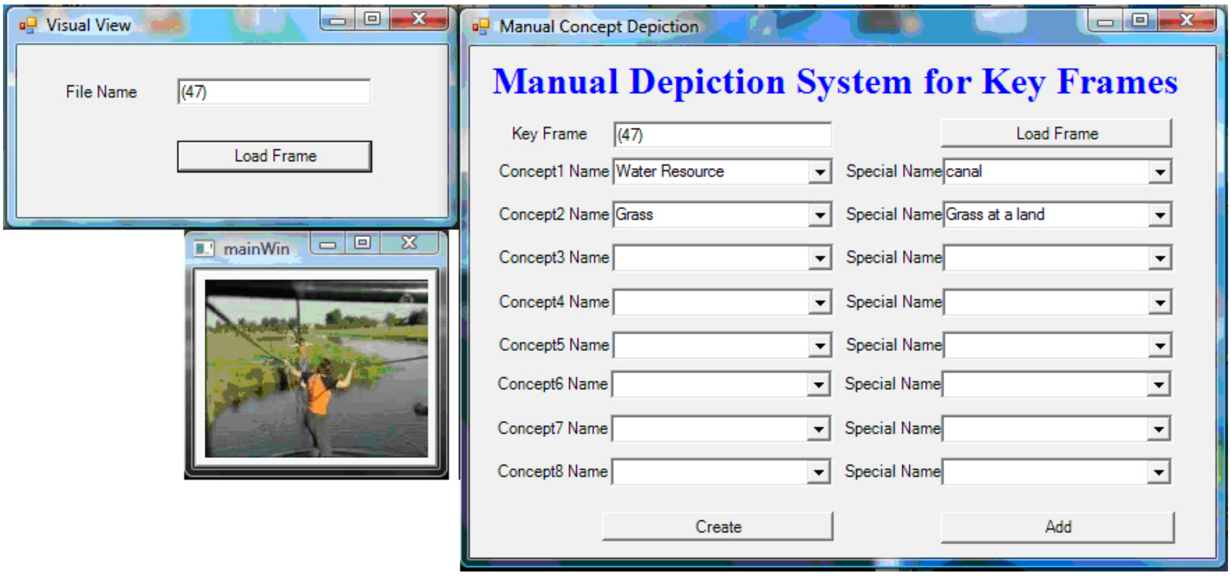
□ Section 13 □ Section 14 □ Section 15 □ Section 16

ALL ABOVE FIELDS SHOULD BE FILLED

Fig. 8: Training system interface

The training interface shown in Fig. 8 enables us to utilise exact frames, and regions in video frames. The selected 2000 video clips have been checked against their TRECVID 2007 development data depictions and our manual depiction system was used to further format it according to the experimental score validation. The manual depiction was done according to a special depiction schema which contains available visual concepts maximum up to 8. Then each test video clip was manually depicted with its visual concepts. For this purpose our manual key frame depiction interface was used. The manual depiction system interface with input fields is shown in Fig. 9.

There is an automated process used to validate the system generated depictions against manual depictions for the use of evaluation. Finally the automated key frame shot depictions generated by experimental systems were compared with the manual depictions using a comparison program. The correct detections, misses and false classifications were counted against each concept. The experiments were conducted with three experimental setups namely, Dominant Colour Descriptor without optional fields, CDPC with Bhattacharyya classifier, and Dominant Colour Descriptor with spatial coherency option.



Manual Depiction System for Key Frames

File Name: (47)

Key Frame: (47)

Concept1 Name: Water Resource — Special Name: canal

Concept2 Name: Grass — Special Name: Grass at a land

Fig. 9: Manual depiction interface

8.0 EVALUATION CRITERIA AND RESULTS

In the field of information retrieval, the precision and recall are the two important measures used for the evaluation. Let *R* be the set of relevant camera shots, i.e. camera shots containing the specific semantic concept one is looking for. Let *A* denote the answer set, i.e. the number of camera shots that are retrieved by the classifier.

Then the precision and the recall power of the system can be defined as follows. According to the above defined sets of *R* and *A*, the following equations can be derived.

$$Hits = R \cap A \tag{21}$$

$$Misses = R \cap A' \tag{22}$$

$$False\ Positives = R' \cap A \tag{23}$$

$$precision\ (p) = \frac{|Hits|}{|A|}. \tag{24}$$

$$recall\ (r) = \frac{|Hits|}{R}. \tag{25}$$

The *F1 score* (also F score or F-measure) is a measure of a test's accuracy. It considers both the precision *p* and the recall *r* of the test to compute the score. The *F1 score* can be interpreted as a weighted average of the precision and recall, where an *F1 score* reaches its best value at 1 and worst score at 0.

$$F1 = 2\frac{(p.r)}{(p+r)}. \tag{26}$$

The positive to negative concept distribution among the test video clips is listed in Table 1.

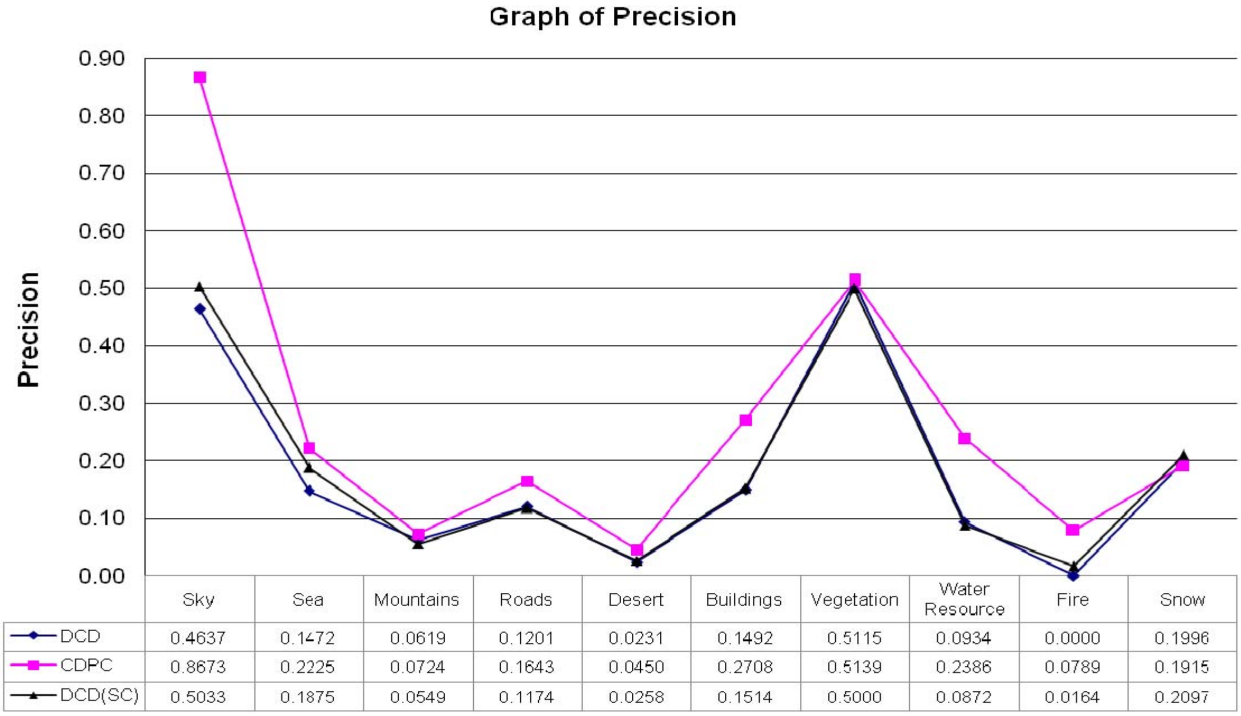
Table 1: Positive to negative ratios of concepts in the test collection

Sky	Sea	Mountains	Roads	Desert	Buildings	Vegetation	Water Resource	Fire	Snow
0.3199	0.0781	0.0235	0.0735	0.0070	0.0764	0.2452	0.0444	0.0040	0.0701

The test collection comprised of a highly unbalanced video clip collections. Since TRECVID has created its videos collections from different sources such as news magazines, science news, news reports, documentaries, educational programs, and archival videos, the visual quality of the collections is highly varied within the subjects

The graph in Fig. 10 shows the precision *(p)* values exhibited by three experimental setups namely, DCD, CDPC with Bhattacharyya (CDPC), and DCD with spatial coherency option (DCD(SC)), for ten irregular shapes based visual concepts.

Fig. 10 clearly illustrates how CDPC experimental precisions stand over other DCD results. The gap between precision values for concepts of sky, building, water resources and fire shows a higher margin. The CDPC experimental precision is only lower to DCD results in the concept of snow with very little margin. The precision values of DCD(SC) has shown a slight improvement compared to DCD.

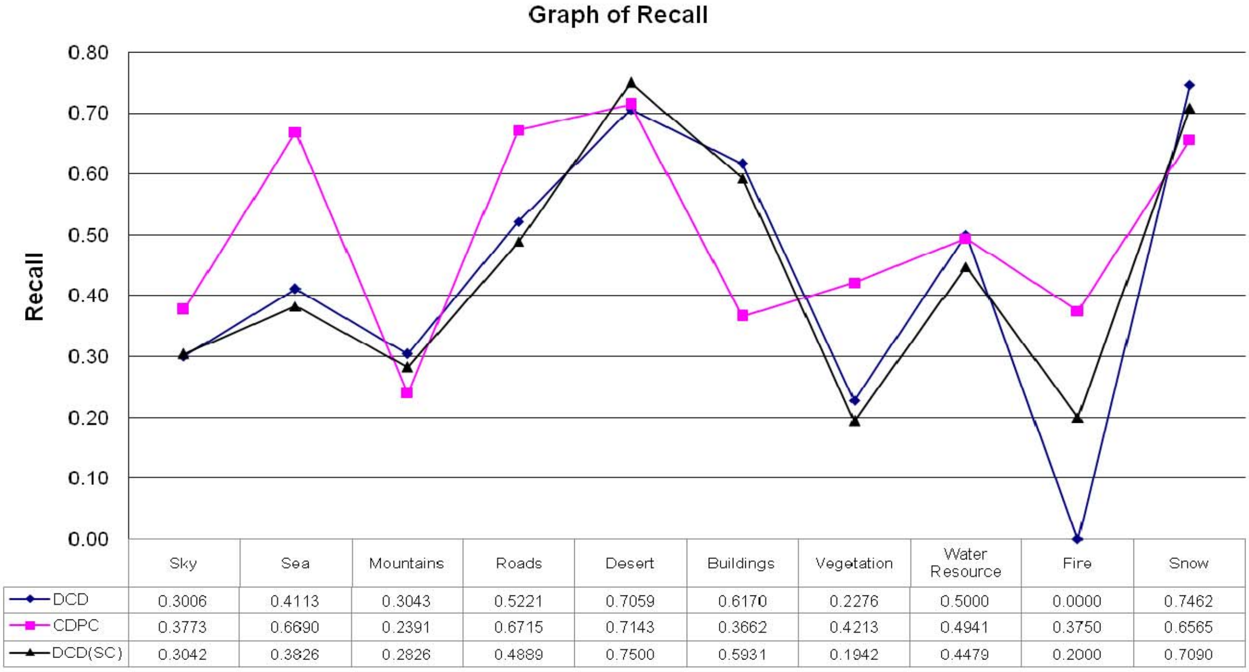


	Sky	Sea	Mountains	Roads	Desert	Buildings	Vegetation	Water Resource	Fire	Snow
DCD	0.4637	0.1472	0.0619	0.1201	0.0231	0.1492	0.5115	0.0934	0.0000	0.1996
CDPC	0.8673	0.2225	0.0724	0.1643	0.0450	0.2708	0.5139	0.2386	0.0789	0.1915
DCD(SC)	0.5033	0.1875	0.0549	0.1174	0.0258	0.1514	0.5000	0.0872	0.0164	0.2097

Fig. 10: Precision values obtained from three experimental setups for ten irregular shapes based visual concepts

The graph in Fig. 11 shows the recall (r) values exhibited by three experimental setups namely, DCD, CDPC with Bhattacharyya (CDPC), and DCD with spatial coherency option (DCD(SC)), for ten irregular shapes based visual concepts.

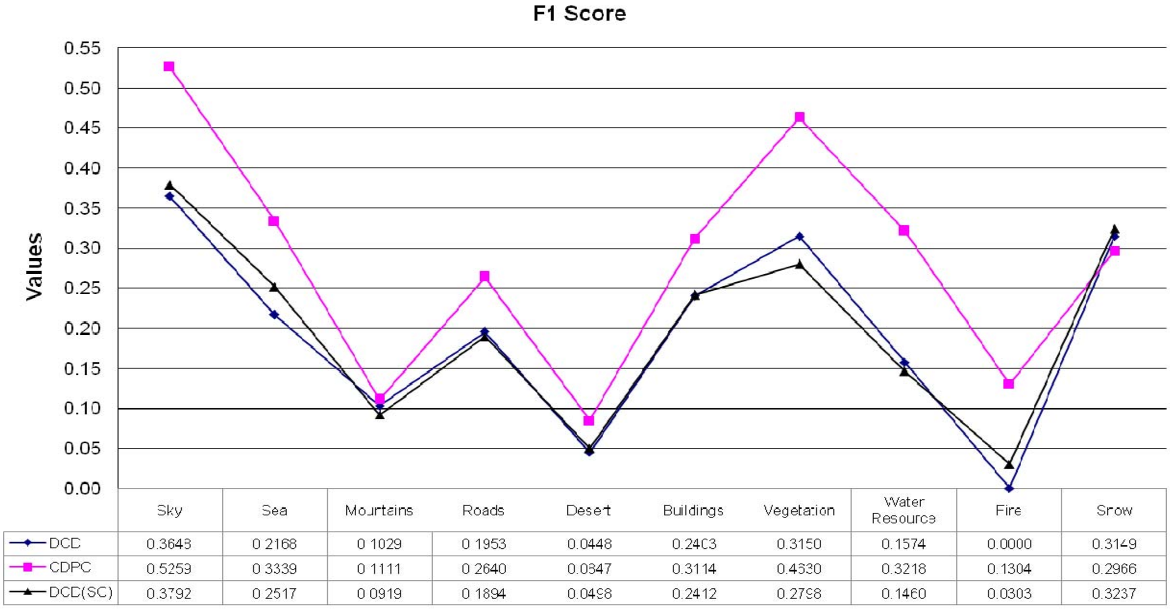
In Fig. 11, CDPC experimental recall results are higher than DCD results for five visual concepts namely sky, sea, roads, vegetation and fire. The average recall power of CDPC for ten concepts lies about 0.498 while DCD(SC) is about 0.435 and DCD is about 0.433. The recall values of DCD(SC) has shown slight variations compared to DCD except for the concept of fire.



	Sky	Sea	Mountains	Roads	Desert	Buildings	Vegetation	Water Resource	Fire	Snow
DCD	0.3006	0.4113	0.3043	0.5221	0.7059	0.6170	0.2276	0.5000	0.0000	0.7462
CDPC	0.3773	0.6690	0.2391	0.6715	0.7143	0.3662	0.4213	0.4941	0.3750	0.6565
DCD(SC)	0.3042	0.3826	0.2826	0.4889	0.7500	0.5931	0.1942	0.4479	0.2000	0.7090

Fig. 11: Recall values obtained from three experimental setups for ten irregular shapes based visual concepts

The graph in Fig. 12 shows the *F1* score results calculated for three experimental setups namely, DCD, CDPC with Bhattacharyya (CDPC), and DCD with spatial coherency option (DCD(SC)), against ten irregular shapes based visual concepts.



	Sky	Sea	Mountains	Roads	Desert	Buildings	Vegetation	Water Resource	Fire	Snow
DCD	0.3643	0.2168	0.1029	0.1953	0.0448	0.2403	0.3150	0.1574	0.0000	0.3149
CDPC	0.5259	0.3339	0.1111	0.2640	0.0847	0.3114	0.4630	0.3218	0.1304	0.2966
DCD(SC)	0.3792	0.2517	0.0919	0.1894	0.0498	0.2412	0.2798	0.1460	0.0303	0.3237

Fig. 12: *F1* score results taken from three experimental setups for ten irregular shapes based visual concepts

Fig. 12 clearly shows how CDPC based experimental *F1* score values are stand over DCD results. The gap between *F1* scores for concepts of sky, vegetation, water resources and fire shows a higher margin. The CDPC based experimental *F1* score is merely lower than DCD results in the concept of snow and that also is with a very slight margin. The *F1* scores of DCD(SC) has shown a very slight variation compared to DCD for all ten concepts.

The Table 2 lists the average results exhibited for all the ten irregular shape based visual concepts by three experimental setups namely, DCD, CDPC with Bhattacharyya (CDPC), and DCD with spatial coherency option (DCD(SC)).

Table 2: Averaged results obtained from three experimental setups for all ten concepts

	DCD	**CDPC**	**DCD(SC)**
p	0.1770	**0.2665**	0.1854
r	0.4335	**0.4984**	0.4352
F1	0.1952	**0.2843**	0.1983

9.0 CONCLUSION

Considering the overall averages exhibited in precision *(p),* recall *(r)* and *F1* measures by CDPC syntactic feature with Bhattacharyya classifier, it is evident that the CDPC based system has performed better over DCD and DCD(SC) systems. The CDPC syntactic colour feature has given a new direction in dealing with complex colour data in videos. Additionally, the CDPC feature space has a number of advantages over MPEG-7 DCD feature space. The compactness of the feature space is enhanced in CDPC with further dimensional reduction compared to DCD feature vector. The efficiency of CDPC based classification has gained advantages from its pattern arrangement compared to DCD distance measure. CDPC feature representation mechanism comes with a compact colour coding system and neighbourhood information. Therefore CDPC enables accurate visual feature classification in digital videos. This is witnessed by higher precisions over DCDs.

The empowerment of the reduction of feature space in CDPC has reduced the complexity of visual concept classification with a low training cost. The experiments were conducted with a highly unbalanced data set, and also the visual quality of the data set was not consistent. However the CDPC based system was able to perform better over MPEG-7 DCD. Even the inclusion of spatial coherency field of DCD was not able to exceed the CDPC based results. The spatial coherency field represents the colour coherency in the selected region. On the other hand CDPC contains information of micro level colour structural arrangements itself, which is similar to texel information.

In the detection of visual concepts with irregular shapes such as sky, vegetation, and other concepts used in this study, colour based visual concept detection and description is dominant. CDPC syntactic feature extends the ability of colour based concept detection and description. CDPC syntactic feature comes with chromatic, spatial and texel information and it has shown greater strength compared to well-known MPEG-7 DCD. It has enriched properties to describe irregular visual scene concepts. CDPC with Bhattacharyya combination has shown higher precisions in the presence of variable visual qualities of video data set. This novel CDPC syntactic feature combined with Bhattacharyya classifier would open up a new direction in video scene concept detection. The inter concept classification interferences and recall power in current results have shown the need to further explore CDPC and Bhattacharyya classification capabilities with the dependant concept of Bhattacharyya coefficient thresholds.

REFERENCES

[1] N. Sebe, “Multimedia Information Retrieval: Promises and Challenges”, *Proceeding of 5th ACM SIGMM International Workshop on Multimedia Information Retrieval,* ACM New, York, 2003.

[2] C.G.M. Snoek, J.C. van Gemert, Th. Gevers, B. Huurnink, D.C. Koelma, M. van Liempt, O. de Rooij, K.E.A. van de Sande, F.J. Seinstra, A.W.M. Smeulders, A.H.C. Thean, C.J. Veenman, M. Worring, “The MediaMill TRECVID 2006 Semantic Video Search Engine”, *Proceedings of TRECVID,* NIST 2006.

[3] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, “Content-Based Image Retrieval at the End of the Early Years”, *IEEE Transactions on pattern analysis and machine intelligence,* vol. 22, no. 12, December 2000, pp. 1349-1380.

[4] A. Mittal, “An Overview of Multimedia Content-Based Retrieval Strategies”, *Informatica,* Vol. 30, No. 3, 2006, pp 347–356.

[5] E. Mr´owka, A. Dorado, W. Pedrycz, E. Izquierdo, “Dimensionality Reduction For Content-Based Image Classification”, *Proceedings of Eighth International Conference on Information Visualisation*, 2004, pp.435-438.

[6] Y. Gao, J. Fan J, “Semantic Image Classification with Hierarchical Feature Subset Selection”, *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, 2005, pp. 135 – 142.

[7] F.E. Badjio, F. Poulet, “Dimension reduction for visual data mining”, *Proceedings of International symposium on applied stochastic models and data analysis*, ASMDA 2005.

[8] H. Bay, A. Ess, T. Tuytelaars, L.V. Gool, “Speeded-Up Robust Features (SURF)”, *Computer Vision and Image Understanding,* Volume 110 , Issue 3, 2008, Elsevier publication, pp. 346-359.

[9] J.M. Martínez (UAM-EPS-GTI, ES), “MPEG-7 Overview (version 10)”, *International Organisation For Standardisation, Coding Of Moving Pictures And Audio,* ISO/IEC JTC1/SC29/WG11 N6828, Palma de Mallorca, 2004.

[10] R. Ewerth, B. Freisleben, “Semi-Supervised Learning for Semantic Video Retrieval”, *Proceedings of ACM International Conference on Image and Video Retrieval*, Amsterdam, Netherlands, 2007, pp. 154-161.

[11] B. S. Manjunath, Jens-Rainer Ohm, V.V. Vasudevan, A. Yamada, “Color and Texture Descriptors”, *IEEE Transactions On Circuits And Systems For Video Technology*, Vol. 11, No. 6, June 2001, pp. 703-715.

[12] R. Datta, J. Li, J.Z. Wang, “Content-Based Image Retrieval - Approaches and Trends of the New Age”, *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, Singapore, 2005, pp 253 – 262.

[13] S. Gao, X. Zhu, Q. Sun, “Exploiting concept association to boost multimedia semantic concept detection”, *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, *ICASSP 2007,* Honolulu, Hawaii, USA, 2007, Vol. 1, pp 981-984.

[14] G. Lavee, L. Khan, B. Thuraisingham, “A framework for a video analysis tool for suspicious event detection”, *Multimedia Tools and Applications*, Vol. 35, Springer Publications, 2007, pp.109–123.

[15] A.K. Heller, Z. Ghahramani, “A Simple Bayesian Framework for Content-Based Image Retrieval”, *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2006, pp. 2110 – 2117.

[16] C. Snoek, et al., “The MediaMill TRECVID 2005 Semantic Video Search Engine”, *Proceedings of the 3rd TRECVID Workshop*, NIST 2005.

[17] N. Sebe, M.S. Lew, “Salient Points for Content Based Retrieval”, *Proceedings of the British Machine Vision Conference 2001*, BMVC 2001, Manchester, UK, British Machine Vision Association 2001, pp 401-410.

[18] K. Mikolajczyk, B. Leibe, B. Schiele, “Local Features for Object Class Recognition”, *Proceedings of the Tenth IEEE International Conference on Computer Vision*, Vol. 2, 2005, pp. 1792 – 1799.

[19] Y.G. Jiang, W.L. Zhao, C.W. Ngo, “Exploring Semantic Concept Using Local Invariant Features”, *Asia-Pacific Workshop on Visual Information Processing*, VIP06, 2006.

[20] H. Shao, Y. Wu, W. Cui, J. Zhang, “Image Retrieval Based on MPEG-7 Dominant Color Descriptor”, *Proceedings of Ninth International Conference for Young Computer Scientists*, IEEE Computer Society, 2008, pp. 753-757.

[21] J. Jiang, Y. Weng, P. Li, “Dominant colour extraction in DCT domain”, *Image and Vision Computing*, Vol 24, Elsevier publication, 2006, pp. 1269–1277.

[22] E. Spyrou, G. Tolias, P. Mylonas, Y. Avrithis, “Concept detection and key frame extraction using a visual thesaurus”, *Multimedia Tools and Applications*, Vol. 41, Springer 2009, pp.337–373.

[23] P. Mylonas, E. Spyrou, Y. Avrithis, S. Kollias, “Using Visual Context and Region Semantics for High-Level Concept Detection”, *IEEE Transactions On Multimedia*, Vol. 11, No. 2, February 2009, pp. 229-243.

[24] W. Surong, C. Liang-Tien, D. Rajan, “Image Retrieval Using Dominant Color Descriptor”, *Proceedings of the International Conference on Imaging Science*, *Systems and Technology*, Las Vegas, Nevada, USA, 2003, pp. 107-110.

[25] B. S. Manjunath, P. Salembier, T. Sikora, *Introduction to MPEG-7: Multimedia content description interface*, WILEY publication, pp. 194-198.

[26] M. Gabbouj, M. Birinci, S. Kiranyaz, “Perceptual Color Descriptor Based on a Spatial Distribution Model: PROXIMITY HISTOGRAMS”, *Proceedings of the International Conference on Multimedia and Computing Systems*, Ouarzazate, Marocco 2009, pp. 144-149.

[27] B.S. Morse, D. Thornton, Q. Xia, J. Uibel, “Image-Based Color Schemes”, *Proceedings of IEEE International Conference on Image Processing*, San Antonio Texas, USA, 2007, pp. 497-500.

[28] G.P.R. Zabi, J. Miller, “Comparing Images Using Color Coherence Vectors”, *Proceedings of ACM Conference on Multimedia*, Boston, Massachusetts, 1996, pp. 65-74.

[29] N.C. Yang, C.M. Kuo, W.H. Chang, T.H. Lee, “A Fast Method for Dominant Color Descriptor with New Similarity Measure”, *Journal of Visual Communication and Image Representation*, Vol. 19, Issue 2, February 2008, pp. 92-105.

[30] A. Bhattacharyya, “On a Measure of Divergence between Two Multinomial Populations”, *Sankhyā: The Indian Journal of Statistics*, Vol. 7, No. 4, Published by Indian Statistical Institute, July 1946, pp. 401-406.

[31] A. Chattopadhyay, A.K. Chattopadhyay C.B. Rao, “Bhattacharyya’s distance measure as a precursor of genetic distance measures”, *J Biosci*., Vol. 29 , No. 2, Indian Academy of Sciences, June 2004, pp. 135–138.

[32] D. Arthu, S. Vassilvitskii, “On the Worst-Case Complexity of the k-means Method”, *Technical Report*, Stanford, 2005.

[33] TRECVID, TREC Video Retrieval at National Institute of Standards and Technology (NIST), http://www-nlpir.nist.gov/projects/tv2007/tv2007.html.

[34] L. Ranathunga, N.A. Abdullah, R, Zainuddin, “Analysis of Video Content in Multi Codec Formats with Compacted Dither Coding”, *Proceedings of Fourth International Conference on Information and Automation for Sustainability, ICIAfS* 2008, IEEExplore, 2008, pp. 48 – 54.