EURASIP Journal on Image and Video Processing manuscript No.
(will be inserted by the editor)

DTCTH: A Discriminative Local Pattern Descriptor for Image Classification

Md. Mostafijur Rahman · Shanto Rahman · Rayhanur Rahman · B. M. Mainul Hossain · Mohammad Shoyaib

Abstract Despite lots of effort being exerted in designing feature descriptor, it is still challenging to achieve acceptable discrimination ability in many image processing applications due to the unavailability of proper generalized descriptor, that itself can capture prominent features for different applications. To address this issue, we propose a Discriminative Ternary Census Transform Histogram (DTCTH) for image representation which uses dynamic thresholds to perceive the key properties of a feature descriptor. The code produced by DTCTH is more stable against intensity fluctuation, and mainly captures the discriminative structural properties of an image by suppressing unnecessary background information. For this purpose, we propose a computationally feasible method which is more generalized and can be used in different applications with reasonable accuracies. To validate the generalizability of DTCTH, we have conducted rigorous experiments on five different applications considering nine benchmark datasets. The experimental results demonstrate that DTCTH performs as high as 28.08% better than the existing state of the art feature descriptors such as GIST, SIFT, HOG, LBP, CLBP, OC-LBP, LGP, LTP, LAID and CENTRIST.

Keywords Discrimination ability · Event classification · Expression recognition · Image classification · Leaf classification · Noise adaptive · Object recognition · Scene classification · Ternary pattern

Md. Mostafijur Rahman, Shanto Rahman, Rayhanur Rahman, B. M. Mainul Hossain and Mohammad Shoyaib are with the Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh
E-mail: bit0312@iit.du.ac.bd, bit0321@iit.du.ac.bd, rayhan@du.ac.bd, raju@du.ac.bd and shoyaib@du.ac.bd

1 Introduction

Image classification has recently gained importance because of its numerous applications in different areas of image processing and computer vision such as texture classification [1–4], object tracking and recognition [5–9], scene classification [5, 7, 10–12], face detection and recognition [13–17], facial expression recognition [17–19], gender classification [17, 20], content based image retrieval [21] and many others. These applications can be incorporated in video surveillance [22], human computer interaction [23], video and image retrieval [24], biometrics [25] and medical imaging [26–28].

Research works in this domain can be grouped into four different categories such as low-level, mid-level, high-level feature representations and classification strategies [29]. Among these, low-level feature description plays a significant role since it is the building block for other steps. Therefore, many feature descriptors have been proposed for low-level feature description. Among these, gradient [10, 30–32] and Local Binary Pattern (LBP) [5, 7, 33] based methods are widely explored, and proved to be successful in different applications. However, in most of the cases, these descriptors solve a particular problem, and fail for general purpose image classification and/or consume high computational cost. To mitigate these problems, in this paper, we intend to develop a computationally low cost general purpose feature descriptor that can perform well in diversified applications. The major challenge is that the real world applications are usually affected by large intra-class and small inter-class variations due to noise, illumination, photometric, scale, rotation, pose and appearance variations [7]. Therefore, it becomes crucial to design a discriminative and robust feature descriptor which will address these issues.

Scale Invariant Feature Transform (SIFT), Histogram of Oriented Gradient (HOG) and GIST are the most commonly used gradient based low-level feature descriptors for image classification [9, 10, 30–32, 34]. Several extensions of SIFT such as Speed Up Robust Features (SURF) [35], Gradient Location and Orientation Histogram (GLOH) [36] and PCA-SIFT [37] have been introduced for improving classification accuracy and/or reducing computational complexity. Besides SIFT, HOG obtains both the properties of SIFT and GLOH [31]. Recently, an extension of HOG, namely Histogram of Second Order Gradient (HSOG) has been proposed to capture curvature information [9]. These descriptors usually use the first derivatives of an image (i.e., gradient direction and magnitude), which can capture local shape properties of the objects.

Gradient-based methods such as SIFT, first generally determine the salient points of an image and then calculate the descriptor on those points. The identification of salient points helps to capture the best discriminative foreground and discard the unnecessary background information. However, the identification of salient point is not directly incorporated to these descriptors. Moreover, in most of the cases these methods do not consider the impact of human visual perception. Further, the gradient-based features often fail to distinguish between two pixels' with same gradients even though those gradients correspond to different local structures [38].

In addition to the gradient based methods, LBP and its extensions such as PRICoLBP [8], DDLBP [39] and OC-LBP [40] have become prominent because of their simplicity and better accuracy [41]. However, LBP based methods that use '0' threshold, have several major drawbacks.

1. Small changes in intensities due to noises in uniform and near uniform regions often lead to wrong LBP codes. For example, in Fig. 1 (b), original intensity '154′ (see Fig. 1 (a)) is changed to '158′, where LBP produces two different patterns (i.e., '11101000′ and '11101100′) though these two textures are similar.
2. LBP based techniques fail to differentiate between the small and large differences in intensities and these also fail to separate the foreground and background which degrades the discriminative ability. For example, differences of center pixel ('170′) and all of its eight neighboring pixels' in Fig. 2 (a) are small and in Fig. 2 (b) are large except one pixel (i.e., '171′), whereas LBP encodes these two textures as same pattern (i.e., '11111111′) which is not desired.

In LBP based methods, all codes are calculated considering the center pixel and hence it can be considered as a background pixel in the local scope. Thus, all its neighbors similar to it should also be considered as background pixel. Since the center pixel is '170′, in Fig. 2 (b), the intensity '171′ should be considered as a background and all other seven neighbors as foreground. However, LBP and most of its variants fail to achieve such discrimination ability.

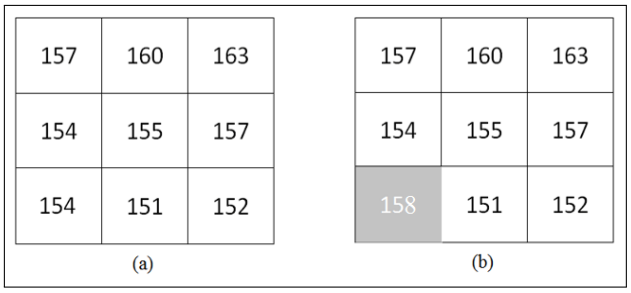


Fig. 1: Noise caused by local intensity fluctuation, (a) original texture, (b) texture changed due to local intensity fluctuation

A similar method to LBP is Census Transform (CT) [4]. Recently, CENTRIST is proposed for scene classification which uses CT of the image pixels' [7]. However, due to the use of static threshold, CENTRIST has similar drawbacks like LBP. In order to address these issues, (i.e., to extract the prominent features from an image and to deal with the presence of different levels of noises) few dynamic threshold based methods are introduced. Local Gradient Pattern (LGP) is one of those which can adapt with local intensity fluctuations by considering mean of the local neighboring differences as a threshold [16]. However, LGP fails to differentiate between a positive and a negative change in the local neighborhoods due to providing same binary code (i.e., '1′) in these two different directions. This problem can be solved by using ternary pattern [2, 3] which cre-

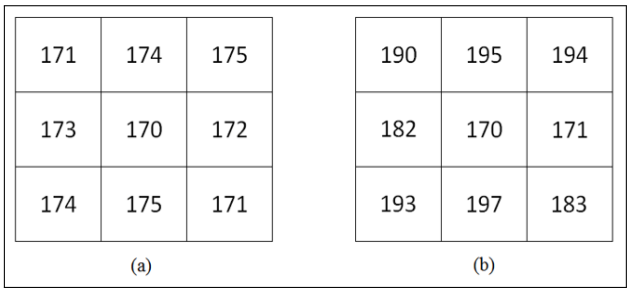


Fig. 2: Example of two different textures which are encoded as same pattern by LBP, (a) small and (b) large differences

ates three patterns instead of two. Among the ternary pattern based techniques, Local Ternary Pattern (LTP) shows resistance to the noises up to a certain level since it assumes that noises in an image usually vary within a fixed threshold ('$\pm 5'$) [2]. However, such a fixed threshold will not work for different types of images [3, 42].

One of the recent attempts of dynamic threshold in ternary pattern has been made in Local Adaptive Image Descriptor (LAID) which considers median of the local neighboring differences to generate the code. However, considering median as a threshold for a general purpose texture description might not be useful in many cases. Because median is determined as the midpoint of data that cannot guarantee the proper separation of significant and insignificant changes. Furthermore, despite the use of median as a threshold, it may have similar drawbacks like LBP, i.e., there might be a case when it will fail to adapt with intensity fluctuation (e.g., produces two different codes '01100011' and '01100111' for the texture in Fig. 1) and cannot discriminate between small and large intensity changes (e.g., produces same code '01100110' for two different textures in Fig. 2).

The incorporation of a non-zero threshold with LBP and its variants usually helps to reduce the effect of noise, suppress the background and highlight the foreground. The benefit of such a threshold can be better realized if it also complies with the Weber's constant [43] because it is not necessary to distinguish the difference in intensity below the Weber's constant, since a person cannot distinguish this change with his/her naked eyes rather it may distract the classifier. The problem in this regard is the choice of threshold that gives the desired result. Hence, the desirable properties of a better threshold is that, it will be able to (i) distinguish foreground and background, (ii) adapt with noise and other lighting conditions, and (iii) consistent with human visual perception.

In this paper, we introduce a new feature descriptor namely Discriminative Ternary Census Transform Histogram (DTCTH) for general purpose image description. The threshold is determined for DTCTH in such a way so that it holds all the desirable properties and can be calculated in linear time. Further, a spatial pyramid representation is used with DTCTH for capturing the global structure of an image. The major contributions of this paper are summarized as follows.

1. We propose a dynamic threshold to produce stable code against intensity fluctuation.
2. The threshold can be calculated in linear time while it preserves all the desirable properties as mentioned above by utilizing only the center pixel. This threshold also helps to separate foreground and background of an image and complies with human visual perception.
3. The proposed DTCTH captures highly discriminative features by suppressing the fine details. Besides, the ternary code is generated to enhance the discrimination ability. We also incorporate a spatial pyramid representation which helps to boost the accuracy.
4. We show the generalizability of DTCTH in case of five different applications such as object, scene, event, leaf and facial expression classification using nine standard datasets.

The rest of the paper is organized as follows. Section 2 briefly discusses the existing state of the art low-level feature descriptors. Section 3 describes the use of these feature descriptors in different applications. The proposed method is described in Section 4. Section 5 presents the rigorous comparative experimental evaluation on five different applications to evaluate the performance of the proposed method. Section 6 concludes the paper with future research scope.

2 Background

A large number of techniques such as GIST, SIFT, HOG, LBP, CLBP, LGP, LTP, LAID and CENTRIST have been proposed for image classification. These techniques capture texture patterns of an image. In this section, a brief description on all of these techniques are highlighted.

2.1 GIST

GIST descriptor is initially proposed in [10] where a low dimensional representation of the scene is developed. The authors propose a set of perceptual dimensions (e.g., naturalness, openness, roughness, expansion, ruggedness) that represent the dominant spatial structure of a scene. The image is divided into small grids (e.g., 4×4 pixels'), for which orientation histograms are extracted using 32 different Gabor filters at 4 scales and 8 orientations. Then the feature values within each grid are averaged. The final GIST descriptor is represented by combining the 16 averaged values of all scale and orientations, which results in $16 \times 32 = 512$ dimensions.

2.2 Scale Invariant Feature Transform (SIFT)

Lowe et al. propose SIFT descriptor which consists of four major steps such as, scale-space peak selection, keypoint localization, orientation assignment and keypoint descriptor [30]. Firstly, potential interest points

are identified in image over scale and space. This is implemented by constructing a Gaussian pyramid and searching for local peaks in a series of Difference-of-Gaussian (DoG) images. Secondly, keypoints are localized to sub-pixel accurately by eliminating inconsistencies. Thirdly, the dominant orientations for each keypoint are identified based on the local image patch. Finally, a local image descriptor is produced for each keypoint, using the image gradients in the local neighborhood. In the representation of the descriptor, gradient locations are quantized into small location grids (e.g., 4×4 pixels'), and the gradient directions are quantized into several (e.g., 8) orientations. SIFT descriptor is represented by combining histograms from all these small location grids (e.g., $4 \times 4 \times 8 = 128$ dimensions). To obtain illumination invariance, the descriptor is normalized by the square root of the sum of squared components.

2.3 Histogram of Oriented Gradient (HOG)

Dalal and Triggs introduce HOG descriptor which takes weighted votes depending on the gradient L2-norm for an orientated histogram channel [31]. HOG descriptor consists of several steps. The image is divided into small connected regions (e.g., 8×8 pixels') named as cells, and a histogram of gradient orientations is computed (e.g., using 1D centered derivative mask $[-1, 0, +1]$) for the pixels' within each cell. Each cell is quantized into angular bins based on the gradient orientation. pixels' in each cell are used as a weighted gradient to the corresponding angular bin. The histogram frequencies are also normalized using L2-norm to adapt with the variation of illumination. The final HOG descriptor is represented by combining these histograms.

2.4 Local Binary Pattern (LBP)

Ojala et al. first explore original LBP operator which thresholds $n \times n$ (e.g., 3×3) neighborhood of every pixel of an image with the center pixel value and considers the result as a binary number [1]. Each of the image pixel is then labeled with the corresponding decimal value of that binary number. The basic LBP is calculated using Equation 1.

$$LBP_{n,r}(x_c, y_c) = \sum_{l=0}^{n-1} q(p_l - p_c)2^l, q(d) = \begin{cases} 1, & if\ d \geq 0 \\ 0, & otherwise \end{cases} \quad (1)$$

Here, n and r are the total number and the radius of the neighboring pixels'. (x_c, y_c) is the co-ordinate of the center pixel c, p_l and p_c are the intensities of the l^{th} neighboring and the center pixel (c) respectively. d is the difference between the neighboring and the center pixel. LBP codes can represent spatial micro-structures such as edge, corner and line-end. Fig. 3 presents some of these patterns.

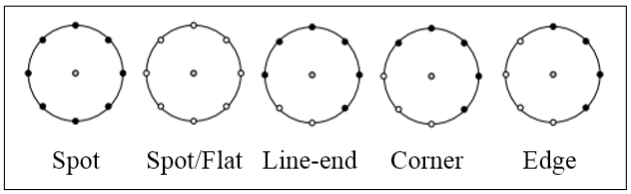


Fig. 3: Example of micro-structures encoded by LBP based methods

LBP has 256 codes when eight neighbors are considered, which can be reduced to 59 codes by taking uniform patterns. The uniform patterns are calculated by Equation 2.

$$U(LBP_{n,r}(x_c, y_c)) = \mid q(p_{n-1} - p_c) - q(p_0 - p_c) \mid + \sum_{l=1}^{n-1} \mid q(p_l - p_c) - q(p_{l-1} - p_c) \mid \quad (2)$$

2.5 Completed Local Binary Pattern (CLBP)

Guo et al. [44] propose CLBP which consists of three components namely CLBP_S, CLBP_M and CLBP_C. CLBP_S considers only the sign value of the differences between a pixel and its neighbors which is exactly same as LBP. CLBP_M uses the magnitudes of the differences between a pixel and its neighbors, and CLBP_C produces code comparing the center pixel's intensity with the average image intensity. CLBP_M is generated following Equation 3.

$$CLBP\_M_{n,r}(x_c, y_c) = \sum_{l=0}^{n-1} q(p_l - p_c)2^l, q(d) = \begin{cases} 1, & if\ d \geq T \\ 0, & otherwise \end{cases} \quad (3)$$

Here, T is the mean of all $|p_l - p_c|$ in the whole image. The CLBP_C is coded as Equation 4.

$$CLBP\_C(x_c, y_c) = q(p_c), q(d) = \begin{cases} 1, & if\ d \geq T_I \\ 0, & otherwise \end{cases} \quad (4)$$

Here, T_I is the average intensity of the whole image. These three operators (i.e., CLBP_C, CLBP_S and CLBP_M) can be combined in two ways. The first way is to build a joint 3D histogram (CLBP_S/M/C), and the second one is to build a 2-D joint histogram by combining CLBP_C with either CLBP_S (i.e., CLBP_S/C) or CLBP_M (i.e., CLBP_M/C). Then this 2-D histogram is converted into a 1-D histogram. Finally, CLBP_M_S/C or CLBP_S_M/C can be generated by concatenating CLBP_M with CLBP_S/C or CLBP_S with CLBP_M/C.

2.6 Local Gradient Pattern (LGP)

LGP is proposed by Jun et al. [16] where $n \times n$ (e.g., 3×3) neighborhood of a pixel is considered, and the neighbor having gradient greater than or equal to the average of gradients of eight neighboring pixels', is set to a binary value of '$1'$, otherwise is assigned a binary value of '$0'$, which is defined by Equation 5.

$$LGP_{n,r}(x_c, y_c) = \sum_{l=0}^{n-1} q(g_l - g_\mu)2^l, q(d) = \begin{cases} 1, & if\ d \geq 0 \\ 0, & otherwise \end{cases} \quad (5)$$

Here, neighboring pixel and mean gradients are calculated as, $g_l = |p_l - p_c|$ and $g_\mu = \frac{1}{n}\sum_{l=0}^{n-1} g_l$ respectively where p_l and p_c are the neighboring and the center pixel's intensities.

2.7 Local Ternary Pattern (LTP)

Inspired by LBP, Tan and Triggs [2] introduce LTP operator. The key difference from LBP is the use of three bits to tackle intensity fluctuation instead of two bits in LBP. Thus LTP produces a ternary code which is calculated using Equation 6.

$$LTP_{n,r}(x_c, y_c) = \sum_{l=0}^{n-1} q(p_l - p_c)3^l, q(d) = \begin{cases} +1, & if\ d \geq\ 5 \\ -1, & if\ d \leq -5 \\ 0, & otherwise \end{cases} \quad (6)$$

Here, (x_c, y_c) is the co-ordinate of the center pixel c. p_c and p_l are the intensities of c and l^{th} neighboring pixels' respectively. To reduce the size of the feature vector, a LTP code is usually split into two binary codes (i.e., upper and lower pattern) and these two types of codes are used for building two histograms separately. Finally, these two histograms are concatenated to represent the feature vector of an image.

2.8 Local Adaptive Image Descriptor (LAID)

LAID is a recently proposed variant of LTP which uses a dynamic threshold to produce a ternary code. LAID operator is defined by Equation 7.

$$LAID_{n,r}(x_c, y_c) = \sum_{l=0}^{n-1} q(p_l - p_c)3^l, q(d) = \begin{cases} +1, & if\ d \geq\ T \\ -1, & if\ d \leq -T \\ 0, & otherwise \end{cases} \quad (7)$$

Here, (x_c, y_c) is the co-ordinate of the center pixel c. p_c and p_l are the intensities of c and l^{th} neighboring pixels' respectively. T is a dynamic threshold which is determined by taking the median of $|p_l - p_c|$. Like LTP, a LAID code is split into two binary codes to reduce the size of the feature vector.

2.9 CENsus TRansform hISTogram (CENTRIST)

CENTRIST is a visual feature descriptor for scene and object classification which performs a Census Transform (CT) of an image and replaces the image with its CT values [7]. CT is a non-parametric local transformation designed for establishing relationships between local patches [4], which is calculated like LBP. CENTRIST does not use interpolation of corner pixels' which is used in LBP. This is the only difference between LBP and CT calculation. The histogram of CT values has been computed to represent the visual descriptor. As CT only encodes the local structures of an image, CENTRIST uses the overlapped spatial pyramid to capture the global structures of an image in large scale. Finally, histograms of all blocks are concatenated to form the feature vector for classification.

3 Literature Review

Till date, SIFT [30] is one of the most successful descriptors in different image processing applications such as scene and object classification. However, one of its major drawbacks is computational cost. Tola et al. propose DAISY descriptor which achieves computational gain by convolving orientation maps using Gaussian kernel [45]. They have used circular regions instead of regular grids where the radius is proportional to the standard deviation of the Gaussian kernel. Comparing different types of spatial pooling scheme, Brown et al. conclude that DAISY style pooling shows better accuracy while keeping lower computational cost [46]. Histogram of second order gradient (HSOG) adopts DAISY pooling, which at first computes a set of first order gradient maps (OGM), then second order gradient is calculated over all OGMs [9], resulted in the increase of both computational cost and accuracy.

SIFT and its variants can capture salient features using key-point descriptors [47] while HOG and its variants use magnitude as weight to deteremine the significance level of saliency in a particular direction. These processes can differentiate background and foreground information implicitly. However, in both cases the computational cost could have been reduced, if the basic descriptor itself were able to identify the salient regions. Besides, most of these methods do not consider human visual perception to distinguish between background and foreground information. Moreover, a gradient based method may fail to differentiate two different textures having the same gradient direction [38].

LBP and its variants [2, 13, 14, 17–20, 41] can capture local microstructures exploring different types of

thresholds. These methods are commonly used for different applications such as face detection [15, 16], human detection [38], object, scene, event [48], face [13, 14], gender [20] and facial expression recognition [18, 49] for their convincing accuracy and lower computational cost. In most of the cases, an image is divided into several blocks where LBP like codes are calculated and then histogram of these codes are calculated for each of these blocks. Finally, these histograms are concatenated to form the final feature vector. A similar but effective variant of this process is described in [18] where Shan et al. use LBP for facial expression recognition adopting boosted SVM. However, the basic LBP only uses sign information. Recently, CLBP is proposed which combines the sign and magnitude to extract more useful information [44]. Because, the combination of sign and magnitude components can provide better clues, which are not evident if only a single component is considered individually [21]. Zhu et al. [40] propose Orthogonal Combination of Local Binary Pattern (OC-LBP) which reduces the dimensionality of the basic LBP from 2^P to $4 \times P$. Due to considering four orthogonal neighbors for each OC-LBP code, this method fails to capture prominent textures even compared to LBP. However, the classification performance is boosted by incorporating bag of features with dictionary learning which increases computational cost. A recent variant of LBP is Local Direction Number Pattern (LDN) [50], which performs well in face and facial expression recognition. LDN encodes the structure of a local neighborhood by analyzing its directional information. Consequently, LDN computes the edge responses in the neighborhood in eight different directions with a compass mask which also introduce extra computational burden.

Recently, Ren et al. have proposed data-driven LBP (DDLBP) for low-level image representation, which is formulated as a point selection problem, that is solved by maximal joint mutual information criterion [39]. This problem is converted into a binary quadratic programming problem and solved efficiently via the branch and bound algorithm. Hussain et al. address that existing local pattern descriptors using hand-specified coding limits those to small spatial supports and coarse gray-level comparisons, and introduce Local Quantized Pattern (LQP) which uses lookup-table based vector quantization to code larger or deeper patterns [51]. LQP inherits some of the flexibility and power of visual-word representations, without sacrificing the speed and simplicity of existing local patterns.

Inspired by the LBP and its variants, several ternary pattern based methods such as LTP [2], NTTP [3] and LAID [42] are also introduced. LAID [42] is a recently explored local ternary pattern for texture classification which uses median of the local neighboring differences as a threshold. However, it may be affected by the non-linear property of median. For example, the median of [0, 1, 1, 2, 3, 4, 17, 18] is 2/3, as a result small differences (e.g., 3, 4) and large differences (e.g., 17, 18) will get the same code which is not expected. Such a scenario (also the opposite scenario [1, 2, 15, 16, 17, 17, 18, 19]) may commonly occur in many applications and thus results in inconsistent code. Hence, LAID may be well for a particular application but might not applicable in general.

Different from LBP, Gabor wavelet feature [52, 53] is one of the major approaches in terms of generality and performance in facial expression recognition. Gu et al. exploit Gabor feature for facial expression recognition which extends the radial encoding strategy for Gabor features based on retinotopic mapping that helps to obtain salient local features for facial expression representation [53]. Another feature descriptor using wavelet theory is Distinctive Efficient Robust Features (DERF) which utilizes exponential scale distribution, exponential grid structure, and circularly symmetric function Difference of Gaussian as convolutional kernel. Thus, it can represent the modeling of response and distribution properties of the parvocellular-projecting ganglion cells (P-GCs) in the human retina. Although Gabor and DERF outperform SIFT, HOG and DAISY, these methods are quite expensive in terms of computational cost.

On top of the basic features, there are few approaches which are used for mid- or high-level image representation [54–57]. Among these, Li et al. propose a high-level image representation named as Object Bank (OB) which describes an image as a scale-invariant response map of a large number of pre-trained generic object detectors [54]. Deformable part-based model (DPM), introduced by Pandey and Lazebnik uses latent SVM for classifying object and scene categories [55]. Besides these, Yang et al. propose spatial pyramid co-occurrence (SPCK++), which calculates spatial co-occurrences of visual words in a hierarchical spatial partitioning [56]. SPCK++ captures both the absolute and relative structure of an image by combining local co-occurrences with global partitioning. Image-To-Class (I2C) distance is first used in NBNN [58] for image classification, which needs higher computational cost for nearest neighbor search in the testing phase. Recently, Wang et al. improve the discrimination of I2C distance especially for small number of local features by learning per-class Mahalanobis metrics [57].

For high-level representation, sparse coding based approaches have shown better performance in image classification which usually adopt SIFT for low-level

feature extraction. One of the first successful techniques is ScSPM [59] which uses sparse coding instead of vector quantization of SIFT descriptors. This technique adopts spatial max pooling (MP) of ScSPM features in regular SIFT grids for final feature representation. ScSPM performs better than both linear SPM kernel (LSPM) on histograms and traditional nonlinear SPM kernels with linear SVM (LSVM) because the pooling of sparse codes quantizes only the essential features which is linearly separable by SVM. However, ScSPM solves L1-norm optimization problem which is computationally expensive [60]. Moreover, it is non-consistent to encode similar descriptors [60, 61]. Several modifications have been proposed for these problems [60–62]. For instance, Wang et al. propose a modification of ScSPM by considering Locality constraints in Linear Coding (LLC) to project each descriptor into its local-coordinate system where projected coordinates are amalgamated by MP [61]. Moreover, ScSPM, LLC and most of the other sparse coding based methods suffer from a severe drawback, which is the quantization of similar local features into different visual words [62]. To mitigate this problem, Oliveira et al. introduce Sparse Spatial Coding (SSC) for image classification which combines a sparse coding dictionary learning, a spatial constraint coding and an online classification stage [62]. The authors represent the final feature vector by adopting MP in SSC features. Most of the sparse coding techniques [59, 61] are adopted on local features independently which consider the global similarity by constraint sparsity. However, dense local features share some local contextual information which is discarded by the existing sparse coding based techniques, and become less reliable when adopting spatial pooling [60]. To address this problem, a Locality-constrained and Spatially Regularized (LCSR) coding is proposed by considering local spatial context of an image into the usual coding strategies which preserves locality constraints both in the feature space and the spatial domain of the image [60]. The information loss in the feature quantization through pooling is still found though several coding methods are introduced to address this problem. Wang et al. use Linear Distance Coding (LDC) to alleviate this problem, which is a complementary technique to the traditional sparse coding schemes [63]. In their approach, local features of an image are transformed into discriminative distance vectors and then encodes these distance vectors into sparse codes to capture the salient features of the image.

Motivated by the sparse coding based approaches, Gao et al. propose kernel sparse representation for image classification which performs sparse coding in a high dimensional feature space mapped by implicit mapping function [64]. Afterwards, by combining these features with SPM, the authors propose Kernel Sparse Representation Spatial Pyramid Matching (KSRSPM). Besides this approach, recently Gao et al. [65] explore another sparse coding based approach (LScSPM) by considering the instable sparse code produced by different sparse coding techniques [59, 61]. The authors use Laplacian sparse coding framework to address this issue. To reduce the high computational cost of dense kernel descriptors, Efficient Match Kernel (EMK) is introduced which maps local features to a low dimensional feature space, and average the resulting vectors to form a set-level feature [66].

Apart from sparse coding based methods, several approaches use soft-assignment coding [12, 67]. For example, Gemert et al. introduce soft-assignment of codewords using kernel density estimation which assigns a local features to all the visual codewords [12]. Comparing with other existing coding scheme, soft-assignment coding is simple and has low computational cost. However, the major drawback is that it cannot produce comparable result with other coding schemes [67]. Liu et al. address that the inferiority of soft-assignment coding is because of its negligence to the underlying manifold structure of local features, and propose a localized soft-assignment coding (LSA) [67]. They use mix-order max pooling (MMP) instead of general MP which helps to boost the performance.

Along with the aforementioned supervised learning techniques, several unsupervised learning techniques are also used in computer vision. For example, Bosch et al. [34] introduce a semi-supervised learning (SP-pLSA) by combining the unsupervised probabilistic latent semantic analysis (pLSA) [68] and a discriminative classifier for image classification. Here, pLSA is applied to the images represented by the frequency of visual-words where color SIFT is used as a basic descriptor. Recently, deep learning based unsupervised technique of feature learning is adopted that does not require manual intervention. This approach has gained popularity because of its better accuracy. Using multiple levels of representation and abstraction, it helps a machine to understand about data (e.g., images, sound and text) more accurately. In deep learning frameworks, first unsupervised feature learning is performed on a large set of image dataset and the weights of the deep network is adjusted. Eventually, a model is built that can later be used to solve a particular problem which is known as fine tuning. Among the existing popular models, AlexNet [69], Places-CNN [70] and VGG_S [71] are widely used because they cover diversified applications. Despite the gain of popularity of deep learning, it is very computation intensive, requires expensive hard-

ware and large set of training data. Furthermore, a well defined network structure is also required to solve a particular set of problems though it is still challenging, and usually fix-up empirically.

The mid- or high-level feature representation aim to capture strong spatial layouts, encode salient textures and make those working with linear classifier [55, 59, 61]. To achieve the aforementioned properties, these methods incorporate different steps such as generative part models [58, 72], discriminative codebook learning [67, 73], sparse coding [59, 61, 65] and/or spatial pooling [61]. The incorporation of these steps lead to increase in computational cost. However, if it is possible to incorporate these issues to the basic feature descriptor, it may reduce the huge computational cost of the mid-/high-level representation.

Apart from these levels of representations (low, mid and high), classifiers also play an important role in classification accuracies. Such as, SVM with different kernels (e.g., linear, polynomial and RBF kernel) are used for classification in various applications [7, 17, 18, 50]. In general, although RBF kernel produces better results in many applications, its computational cost is high. A fast and effective classification is thus necessary which can be achieved in two ways such as by selecting relevant features where nonlinear relationship of features is already incorporated and then use LSVM, or by introducing a low cost nonlinear kernel of SVM. Maji et al. [74] introduce a fast non-linear kernel of SVM namely histogram intersection kernel which achieves better classification accuracy in many applications [75]. Zhang et al. propose a hybrid classification technique (SVM-KNN) which selects features using k nearest neighbors [76] and classify using DAGSVM [77] classifier. SVM-KNN has low computational cost, and performs well when the test image is similar to one of the training images. However, this technique fails to generalize much beyond the labeled images because of calculating Image-to-Image distance. Recently, to perform fast and better classification, Jianxin Wu [78] introduces PmSVM, which solves a dual SVM formulation using a coordinate descent approach. PmSVM approximates the gradient using polynomial regression instead of the kernel function and feature mapping.

From the above discussion, it can be seen that most of the existing techniques attempt to capture the salient textures that are stable against different lighting conditions, noise, intensity fluctuation and clearly represent necessary foreground information. For this purpose, these approaches either include preprocessing such as keypoint identification before generating descriptor or postprocessing such as different costly high-level representations. However, the computational cost can be reduced if it is possible to identify the prominent features using only the basic low cost descriptor. Therefore, it is necessary to incorporate such a mechanism with the low-level descriptor. In case of LBP based methods, the choice of a proper threshold can ensure this desirable property of a low-level descriptor.

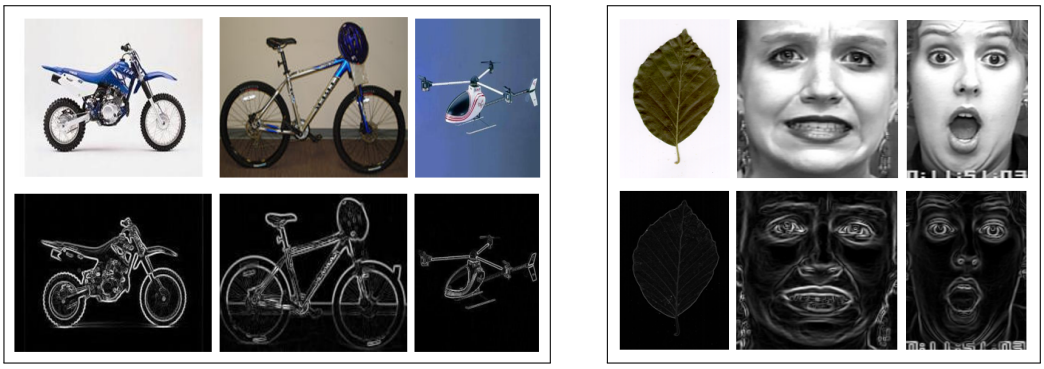
4 Proposed Method

In this paper, we propose a new feature descriptor named as Discriminative Ternary Census Transform Histogram (DTCTH) for image representation which holds most of the key properties of a feature descriptor. The overall process of the descriptor construction is described in the following subsections.

4.1 Desired Properties

A feature descriptor for image classification should have the following essential properties.

1. **Discrimination Ability**: A feature descriptor should have higher discrimination ability. If it has the capability to encode only the class specific information by suppressing the unnecessary background, it will perform well in image classification. Fig. 4 presents several images with corresponding Sobel images from different object, scene and expression classes. All of these images contain respective class specific information which is clearly visualized from their Sobel images. This class specific information needs to be encoded for better image classification. Therefore, our goal is to encode only this class specific information by discarding the unnecessary background details.
2. **Illumination Invariance**: A good feature descriptor should be able to adapt with illumination changes because illumination of same images can vary due to different reasons. Among the existing low-level feature descriptors, CENTRIST based methods have this property and if we follow the basic CENTRIST structure, our proposed descriptor will have the same property.
3. **Generalizability**: It is expected that a descriptor has reasonable accuracy for different types of applications. This can be achieved when a descriptor is capable to encode class specific features and suppress unnecessary background information for the respective applications. We will design our descriptor such a way that it will have this property.
4. **Incorporation of Visual Perception**: In general, a person cannot distinguish a change in an image if



(a) (b)

(c)

Fig. 4: Sample images with corresponding Sobel images from different categories of object, scene and expression (first row original images and second row Sobel images), (a) object classes, (b) leaf and expression classes, (c) scene classes

it is below the Weber's constant [43]. So, it is reasonable to consider that the change below this constant is not necessary to capture for machine vision. Thus a descriptor should have the capability to capture only those changes that inline with human vision.

5. **Stable Code**: Producing stable code (i.e., same code) against intensity fluctuation is another essential property for a feature descriptor. It is obvious that intensity of an image might be changed for several reasons. Let $\delta = p_l - p_c$, where p_c is the intensity of a target pixel c and p_l is the intensity of its l^{th} neighbor. If the difference of intensities, $\mid \delta \mid$, of the two pixel is large, those two pixels' should be considered differently and vice versa. Thus, the range of δ has to be set in such a way so that the two pixels' get the same or different codes in two different situations. At this point, we define two terms *certain* and *uncertain* state for a code (C) using Equation 8.

$$C = \begin{cases} certain\ state, & if \ \mid \delta \mid \geq T \\ uncertain\ state, & otherwise \end{cases} \tag{8}$$

Here, T is a threshold that might be static or dynamic. Defining *certain* and *uncertain* states have several advantageous properties. For example, in this case we can achieve discriminative and stable code because of considering the certain and uncertain states separately. Apart from that, we can get three groups (G) of codes using Equation 9. Group one ($g1$) and group three ($g3$) belongs to certain state, and group two ($g2$) remains in uncertain state.

$$G = \begin{cases} g1, & if \ \delta \geq T \\ g2, & if \ -T < \delta < T \\ g3, & if \ \delta \leq -T \end{cases} \tag{9}$$

Again T should be dynamic because a static threshold might fail in case of different types of images.

4.2 Discriminative Ternary Census Transform Histogram (DTCTH)

The overall process of DTCTH calculation is shown in Fig. 5. For producing different codes for certain and

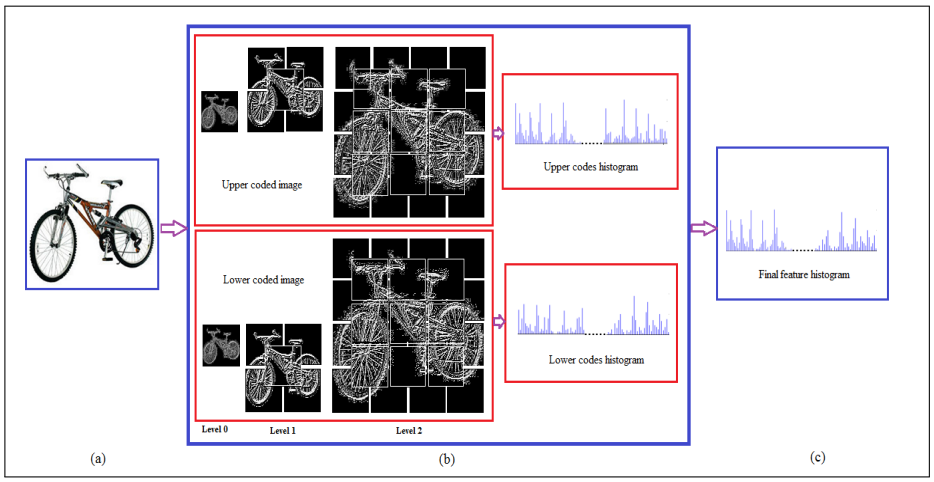


Fig. 5: Overall process of DTCTH calculation, (a) input image, (b) DTCTH codes in overlapping SPM, and (c) final feature histogram

uncertain changes of intensities in an image, we consider ternary coding scheme, namely Discriminative Census Transform (DCT) which is calculated using Equation 10.

$$DCT_{n,r}(x_c, y_c) = \sum_{l=0}^{n-1} q(p_l - p_c)3^l, q(d) = \begin{cases} +1, & if\ \ d \geq T \\ -1, & if\ \ d \leq -T \\ 0, & otherwise \end{cases} \quad (10)$$

Here, T is a dynamic threshold, n and r are the total number of and the radius of the neighboring pixels' respectively, (x_c, y_c) is the coordinate of the center pixel. p_c and p_l are the intensities of the center pixel c and l^{th} neighboring pixel. For simplicity and computational efficiency, the ternary pattern is divided into two census transformed images namely upper (DCT_UP) and lower (DCT_LP) pattern which are calculated using Equation 11 and 12. Fig. 6 shows a pictorial example of DCT calculation. Afterwards, two separate histograms such as H_DCTUP and H_DCTLP of these two binary patterns are calculated using Equation 13 and 14. The final feature vector is represented by concatenating these histograms.

$$DCT\_UP_{n,r}(x_c, y_c) = \sum_{l=0}^{n-1} q(p_l - p_c)2^l,\ q(d) = \begin{cases} 1, & if\ \ d \geq T \\ 0, & otherwise \end{cases} \quad (11)$$

$$DCT\_LP_{n,r}(x_c, y_c) = \sum_{l=0}^{n-1} q(p_l - p_c)2^l,\ q(d) = \begin{cases} 1, & if\ \ d \leq -T \\ 0, & otherwise \end{cases} \quad (12)$$

$$H\_DCTUP^k = \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} \delta^k_{DCT\_UP_{n,r}(i,j)},\ \delta^k_p = \begin{cases} 1, & if\ p = k \\ 0, & otherwise \end{cases} \quad (13)$$

$$H\_DCTLP^k = \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} \delta^k_{DCT\_LP_{n,r}(i,j)},\ \delta^k_p = \begin{cases} 1, & if\ p = k \\ 0, & otherwise \end{cases} \quad (14)$$

Here, $DCT\_UP_{n,r}(i,j)$ and $DCT\_LP_{n,r}(i,j)$ are the upper and lower DCT codes of coordinate $(i,\ j)$. k is the k^{th} bin of the histogram. h and w are the height and width of the image block. Kronecker delta δ^k_p is a piecewise function of p and k. As DCT encodes only local micro-structures of spatial location, Spatial Pyramid Representation (SPM) used in CENTRIST is adopted to capture the global structures of an image.

4.3 Determining the Value of T

Let us consider, we have eight δ values for a pixel and want to partition these values into two groups using a threshold such that the variance is maximized between the groups and is minimized within the group. In this partition, the group with lower and higher values can be considered as background and foreground respectively, in the local scope. Using Jenk's Natural Breaks optimization method [79, 80], a solution of this problem can be found. However, such an optimization is very time consuming as we have to calculate this value for each pixel of an image to generate the respective code. Furthermore, Jenk's Natural Breaks optimization is not designed to comply with Weber's constant. Whereas, the combination of these two is expected to increase the accuracy. But, this combination is not trivial in a simple operation. Under these circumstances, after exhaustive empirical analysis (on 10^9 samples), we have come up with the idea to set the value of T to the square root of the center pixel in a local binary pattern. We have found that such a choice of T brings about the closest possible similarity which is around 84.90%, to the aforementioned optimization problem considering Weber's constant. Thus, we can conclude that taking the square

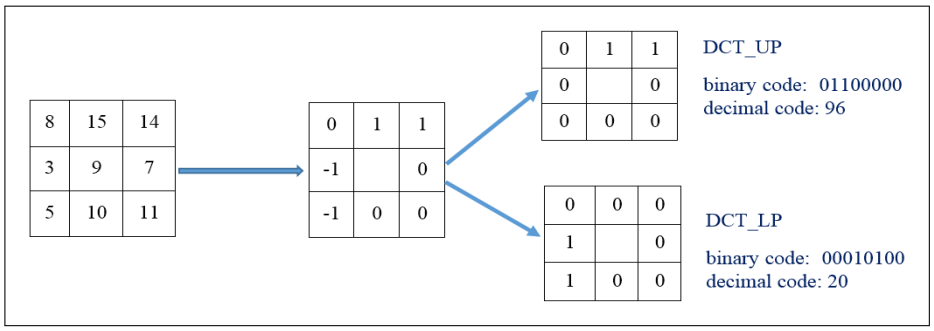


Fig. 6: An example of DCT calculation ($T = 3$)

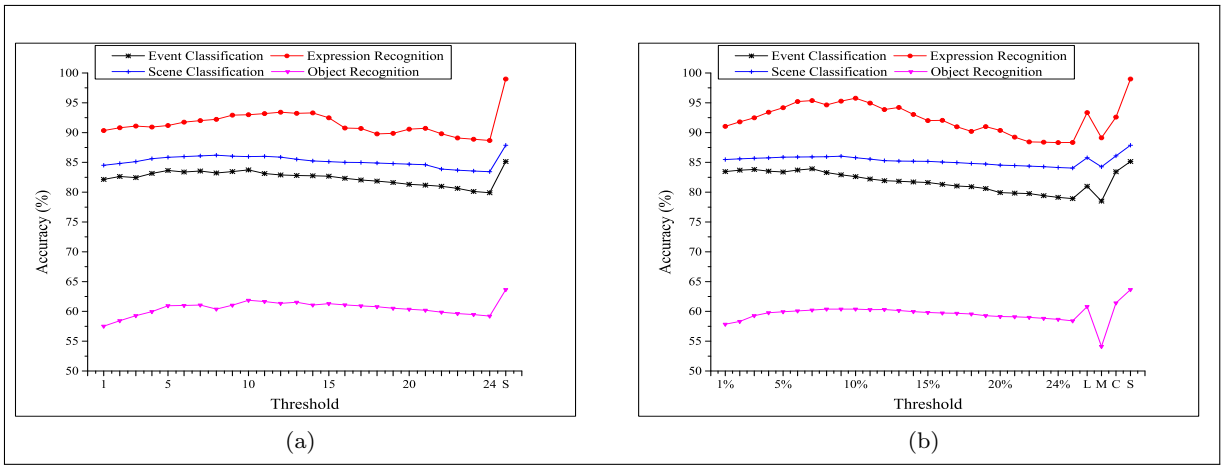


Fig. 7: Demonstration of object recognition, scene classification, event classification and expression recognition accuracy using different thresholds, (a) accuracy of different constant thresholds, (b) accuracy of different dynamic thresholds (i.e., percentage of the center pixel as threshold). All the results are generated considering ternary pattern in CENTRIST framework. Noteworthy, L (threshold used in LGP [16]), M (median threshold used in LAID [42]), S (SQRT of center pixel) and C (cube root of center pixel)

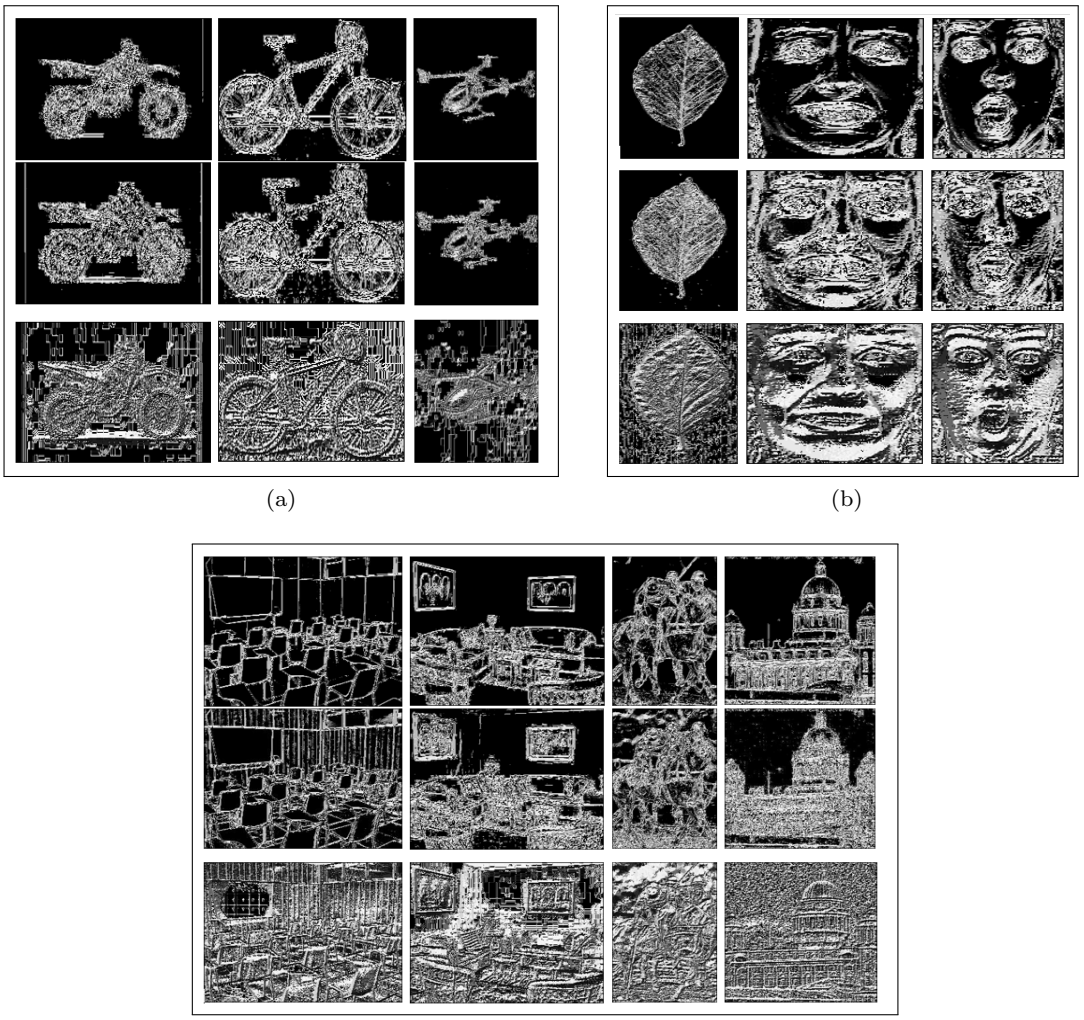
root of the center pixel is a very close approximation of the desired threshold with much low computational cost.

For validating the value of T, we have performed rigorous experiment on four different applications with four datasets using different values of T. The dataset includes Caltech-101 [72] (102 classes and 9,145 images) for object classification, UIUC Sport Event [33] (8 classes and 1,586 images) for event classification, OT scene [10] (8 classes and 2,688 images) for scene classification and Cohn Kanade [81] (6 classes and 960 images) for expression recognition. We consider both the fixed and dynamic thresholds to determine the value of T. From the experiments, we observe that the accuracy is decreased for the values greater than 20, and hence we consider the values upto 25, both in fixed and dynamic cases. The mean and median of the differences among the neighboring pixels' and the center pixel, SQRT, and cube root of the center pixel are also considered. Fig. 7 (a) shows the accuracy of different fixed thresholds and square root threshold, and Fig. 7 (b) illustrates the accuracy of different dynamic thresholds, as mentioned earlier. From Fig. 7, it is found that T is defined as SQRT of center pixel, performs best for all applications. Using *McNemar's* test, we observe that the proposed SQRT threshold resulted in significantly fewer mis-classifications than other thresholds (maximum p value, $P = 0.001$ and minimum p value, $P = 3.83932E - 28$).

4.4 Properties of DTCTH

DTCTH encodes micro-structures such as line, edge and corners which are stable against intensity fluctuation and monotonic illumination variation. Some of these properties are described in the following.

DTCTH captures more relevant part of an image that are necessary for recognizing an object/scene. To



(a) (b) (c)

Fig. 8: Coded image by DTCTH, LTP and CENTRIST for corresponding images in Fig. 4, (a) Object classes, (b) Leaf and expression classes, (c) Scene classes (First row - DTCTH coded image, second row - LTP coded image, and third row - CENTRIST coded image)

understand this, let us consider Fig. 4 which includes examples from three different applications. Now, if we only have the Sobel images where all fine details are suppressed and only the class specific information is retained then it will help a classifier to achieve better accuracy. Likewise, if we analyze the images in Fig. 8, we can easily find that DTCTH suppresses most of the background keeping the necessary details compared to the others. As the proposed technique have this property, it is more generalized compared to other descriptors.

DTCTH features are more robust to noise, and produce stable code by adapting the intensity fluctuations in local neighborhood. For example, CENTRIST and LBP fail to produce the same code in case of intensity fluctuation (see Fig. 1), whereas DTCTH is successful in this case (i.e., '00000000′). Furthermore, we add white Gaussian noise to the original images as shown in Fig. 9 to test the noise robustness of DTCTH. Now, if we compare the coded images with or without noises for DTCTH and LBP, we can easily find that DTCTH is more robust to noise and thus can capture the face specific feature by eliminating the details.

Besides these, for certain intensity changes in positive and negative directions, DTCTH produces two different codes for these two directions which is desired, because from this type of representation, we can get more detailed information about the local microstructure of an image. For example, in Fig. 10, DTCTH produces three different codes for aforementioned three

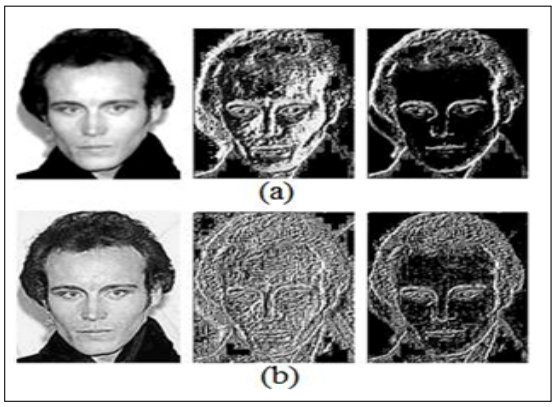


Fig. 9: Noise resistance of CENTRIST vs. DTCTH, (a) image without noise, (b) noisy image (middle one is CENTRIST coded image, right one is DTCTH coded image)

80	81	79
71	70	69
61	60	60

Fig. 10: Illustrative example of the certain and uncertain regions in an image

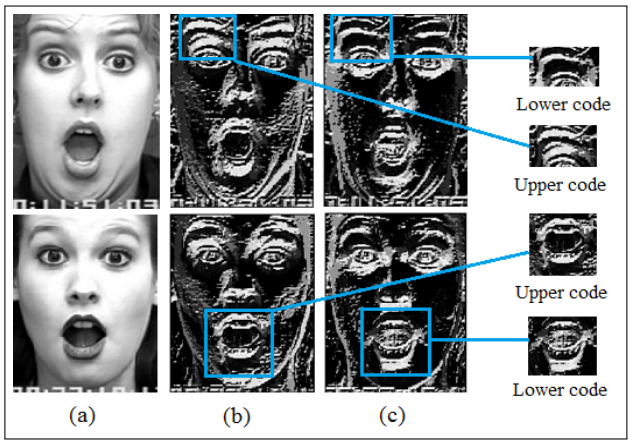


Fig. 11: DTCTH coded image, (a) input image, (b) upper code, (c) lower code

groups of codes following Equation 9 such as uncertain state (i.e., 0 for 71 and 69), intensity changes in positive direction (i.e., 1 for 80, 81 and 79) and negative direction (i.e., -1 for 61, 60 and 60) in certain regions by considering 70 as the center pixel.

To understand the effect of human visual perception in case of DTCTH, we require a reference value to measure the change in intensity. Since, DTCTH uses the center pixel for calculating code, we use the same reference point for measuring Weber's constant. For example, in Fig. 10, the Weber's constant for the neighboring pixels' 71 and 69 with the center pixel 70 is 0.01 which is below the Weber's constant for human visual perception [43]. Hence, these two neighbors are coded as '0'. Similarly, 79 and 61 have the Weber's constant of 0.13, hence these two pixels' are coded as '1' and '-1' due to the changes in two different directions which is expected. From Fig. 11, it is understandable that the smooth regions (e.g., cheek area) are coded similarly and the codes for $+ve$ and $-ve$ directions contain complementary information (e.g., eyebrow and mouth regions).

5 Experimental Evaluation

In this section, we evaluate the performance of DTCTH by comparing with the other state of the art methods over nine datasets that belong to five different applications such as object, scene, event, leaf and facial expression classification. Application wise descriptions of the experiments are discussed in the following.

Table 1 presents the overview of these nine datasets along with the number of training and testing samples used in the experiments, which is also described in the respective datasets. For the experiments, all images are resized to at most 300×300 pixels'. Except the facial expression recognition, the dataset is split into five random partitions and experiments are performed five times. That is, we perform five fold cross validation and report the average accuracies in the respective tables. In case of facial expression recognition, the experiments are run ten times with person independent splits by following the standard protocol, and the average accuracies are reported in the tables. The datasets description, followed by the proper comparison with state of the art methods, are described in details in the following subsections. In this paper, we also provide results of some of the deep learning based techniques for completeness. However, these techniques are not directly comparable to DTCTH.

For implementing DTCTH, few parameters are involved with the basic descriptor (DTCTH) and its classifier (SVM). The major parameters for DTCTH are its radius (r) and its number of neighbours (n). From literature, we have found the best accuracies with reasonable feature vector length are produced using $r = 1$ or $r = 2$ and $n = 8$ in most of the applications [13]. For SVM, different types of kernels such as Linear, RBF, Polynomial, Sigmoid and Histogram Intersection (HI) can be used. For the first four kernels, we use LibSVM

Table 1: Different benchmark datasets with proper training samples

Applications	**Object Classification**		**Event Classification**	**Scene Classification**			**Leaf Classification**	**Facial Expression Classification**	
Databases	Caltech-256	Caltech-101	UIUC Sports Event	OT Scene	Scene 15	Indoor 67	Swedish Leaf	Cohn Kanade (CK)	CK+
Classes	257	102	8	8	15	67	15	6/7	7
Total Samples	30,608	9,145	1,586	2,688	4,485	5,620	1,125	960/1,280	981
Training Images/class	60	30	70	100	100	80	25	Person Independent 10 fold cross-validation	
Test Images/class	Rest	Rest	60	Rest	Rest	20	Rest		

package[1]. To find out how DTCTH behaves with these parameters, we use three datasets namely Caltech 101, UIUC Sports Event and Scene 15. The results with these parameters settings for theses datasets are summerized in Table 2 which show that DTCTH considering 8 neighbors at radius 2 using HI kernel works well in most of the cases.

In this work, we mainly adopt the CENTRIST framework[2], keeping all the parameters' same as described in CENTRIST [7]. Thus for fair comparison, we consider *eight* neighbors at radius *one* from the center pixel like CENTRIST in all the experiments, although consideration of other parameter setting may produce better result. Following CENTRIST, we also avoid corner points interpolation and remove two DCT bins (i.e., 0 and 255) while calculating DCT histograms. Afterwards, we take the square root of DTCTH histogram and perform L1 normalization of those descriptors. For classification, we use SVM classifier with linear kernel ($c = 2^{-5}$, $g = 2^{-7}$) [82] and Histogram Intersection (HI) kernel [74]. We use the aforementioned parameter settings unless otherwise stated. To reflect a brief description of a particular method, we mainly consider the following representation for Tables 3-11. Firstly, we give the basic descriptor name followed by mid- /high-level representation in the parentheses, then the classifier name and publication year.

5.1 Object Classification

We have considered two well-known and most challenging object datasets named as Caltech-101 [72] and Caltech-256 [11] to evaluate the object recognition performance of the proposed descriptor. These two datasets are described below followed by the obtained results from the experiments.

[1] LibSVM- `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[2] CENTRIST Framework- `http://cs.nju.edu.cn/_upload/tpl/00/ed/237/template237/publication.htm`

Table 2: Effect of different SVM and DTCTH parameters on UIUC Sports Event, Caltech 101 and Scene 15 datasets. Here, we consider 70 training and 60 test images for UIUC Sports Event, 30 training and remaining test images for Caltech 101, and 100 training and remaining test images for Scene 15.

Techniques	UIUC Sports Event	Caltech 101	Scene 15
Linear Kernel			
$DTCTH_{8,1}$	85.16±0.96	72.26±1.67	82.66±0.50
$DTCTH_{8,2}$	84.73±1.01	76.08±0.41	82.87±0.49
Polynomial Kernel			
$DTCTH_{8,1}$	83.69±0.97	68.64±0.53	80.92±0.12
$DTCTH_{8,2}$	84.02±1.15	73.21±0.58	82.62±0.56
RBF Kernel			
$DTCTH_{8,1}$	75.74±1.54	58.72±1.16	72.95±0.62
$DTCTH_{8,2}$	75.83±1.17	63.96±1.36	73.73±0.65
Sigmoid Kernel			
$DTCTH_{8,1}$	67.95±1.94	52.59±1.41	68.16±1.88
$DTCTH_{8,2}$	70.47±1.58	56.88±1.31	69.59±1.08
Histogram Intersection Kernel			
$DTCTH_{8,1}$	88.18±0.84	78.56±0.91	83.63±0.21
$DTCTH_{8,2}$	87.75±0.57	80.36±0.24	83.92±0.43

5.1.1 Caltech-101

Caltech-101 contains 9,144 images of 101 categories and an additional background category, making a total number of 102 categories, with significant variance in shape [72]. The number of images per category varies from 31 to 800. As suggested by the original dataset [72] and many other researchers [5–7, 59], we have partitioned the whole dataset into 5, 10, 15, 20, 25 and 30 training images per class and rest for testing to measure the performance unless otherwise stated.

To compute DTCTH code, it only compares its pixel values with a specific value (square root of the center pixel), and performs better than SIFT, DAISY and HSOG techniques. DTCTH achieves 78.56% accurate object classification rate by considering only the low-level feature representation, which demonstrates the improvement of performance over existing state of the art methods such as SSC [62], ScSPM [59], LSPM [59], LLC [61], LCSR [60] and LDC [63]. Even though most of

Table 3: Object classification rate (%) in Caltech-101

Techniques	5	10	15	20	25	30
Places-CNN, 2014 [70]	-	-	-	-	-	65.18
ImageNet-CNN, 2014 [70]	-	-	-	-	-	**87.22**
Hybride-CNN, 2014 [70]	-	-	-	-	-	84.79
dense color SIFT (SP-pLSA) SVM, 2008 [34]	-	-	59.8 (50)*	-	-	67.7 (50)*
SIFT (ML + CORR) KNN, 2008 [83]	-	-	61.00	-	-	69.60
SIFT (ML + PMK) KNN, 2008 [83]	-	-	52.20	-	-	62.10
dense SIFT (KC) SVM with HI, 2008 [12]	-	-	-	-	-	64.14 (50)*
dense SIFT (LSPM + MP) LSVM, 2009 [59]	-	-	53.23	-	-	58.81
dense SIFT (ScSPM + MP) LSVM, 2009 [59]	-	-	67.0	-	-	73.2
dense SIFT (LLC + MP) LSVM, 2010 [61]	51.15	59.77	65.43	67.74	70.16	73.44
dense SIFT (LSA + MMP) LSVM, 2011 [67]	-	-	-	-	-	74.21
dense SIFT (LDC + LLC/LSA + MP) LSVM, 2013 [63]	-	-	-	-	-	74.47
dense SIFT (LCSR + MP) LSVM, 2012 [60]	-	-	-	-	-	73.23
dense color SIFT (GOLD) LSVM, 2015 [84]	-	-	73.39 (at most 50)*	-	-	80.92 (at most 50)*
dense SIFT (SSC + MP) OCL, 2012 [62]	55.64	65.52	69.98	73.99	75.49	77.59
HSOG (LLC + MP) SVM, 2014 [9]	-	-	60.46 (15)*	-	-	67.97 (15)*
CSIFT (LLC + MP) LSVM, 2015 [85]	46.48	56.97	62.09	65.45	68.17	69.18
dense SIFT (BoF) SVM, 2004 [9, 30]	-	-	62.48 (15)*	-	-	69.89 (15)*
$CS-LBP_{2,8,0.01}$ (BoF) SVM, 2009 [9, 86]	-	-	58.5 (15)*	-	-	66.86 (15)*
DAISY (BoF) SVM, 2010 [9, 45]	-	-	58.63 (15)*	-	-	67.01 (15)*
SIFT (SPM) SVM, 2006 [5]	-	-	56.40 (50)*	-	-	64.60
dense SIFT (SPM) SVM, 2007 [11]	44.2	54.5	59.0	63.3	65.8	67.60
dense SIFT + NBNN, 2008 [58]	-	-	65.00 (20)*	-	-	70.4
geometric blur + SVM-KNN, 2006 [76]	46.6	55.8	59.05	62	-	66.23
dense SIFT (BoF) PmSVM-χ^2, 2012 [78]	-	-	72.08 (20)*	-	-	-
dense SIFT (BoF) PmSVM-HI, 2012 [78]	-	-	72.18 (20)*	-	-	-
LGP (SPM) LSVM, 2013	39.86	50.11	57.84	60.03	62.96	66.52
OC-LBP (BoF) LSVM, 2013	47.10	56.34	62.43	64.70	67.63	70.87
LAID (SPM) LSVM, 2013	39.03	48.35	54.11	57.83	60.84	63.87
CLBP_S/M/C (SPM) LSVM, 2010	32.06	40.03	45.59	49.40	52.56	55.35
LTP (SPM) LSVM, 2010	41.04	51.23	59.69	61.17	64.57	67.85
GIST + LSVM, 2001	40.16	47.87	52.5	56.25	58.88	61.70
CENTRIST (SPM) LSVM, 2011	39.46	49.72	55.84	59.47	62.25	65.23
Proposed (DTCTH + LSVM)	46.98	57.00	63.66	65.83	68.69	72.26
Proposed (DTCTH + HI)	56.74	65.97	71.84	74.80	76.85	78.56

* Different number of test images used for the experiment rather than standard settings

these methods use different high-level representations. A recent state of the art method namely Gaussians of Local Descriptors (GOLD) [84] that achieves 80.92% accuracy using 30 training and only 50 testing samples. It uses dense SIFT as a basic descriptor and focuses on high-level representation. The result is comparable when we use DTCTH with $r = 2$. However, their computational cost is much higher compared to us. Colored SIFT (CSIFT) [85] is another recent state of the art low-level descriptor that also uses LLC as a high-level representation. However, this method produces inferior result compared to DTCTH.

Apart from the aforementioned techniques, other well known descriptors such as GIST [10], CENTRIST [7], LTP [2] and LGP [16] are used in different applications and compared with DTCTH. For the sake of fair comparison, the results of CENTRIST, LTP and LGP are generated using same parameter settings that we have used. The result of GIST descriptor is generated using the standard setup, which is 32 Gabor filters in 4 scales and 8 orientations. All of these low-level feature descriptors produce inferior results in comparison with DTCTH (see Table 3). It is observed from this table that PmSVM [78] performs (72.18%) slightly better than DTCTH (71.84%). It is noteworthy that, they have used different classifier than ours, and considered only 20 sample images for testing.

5.1.2 Caltech-256

Caltech-256 is a very challenging dataset which contains 30, 607 images of 256 categories and an additional clutter category [11]. Each class has at least 80 images which show higher variability in object size, location

Table 4: Object classification rate (%) in Caltech-256

Techniques	15	30	45	50	60
Places-CNN, 2014 [70]	-	-	-	-	45.59
ImageNet-CNN, 2014 [70]	-	-	-	-	67.23
Hybride-CNN, 2014 [70]	-	-	-	-	65.06
SIFT (SPM + pLSA) SVM, 2006 [5]	-	34.10	-	-	-
dense SIFT (LSPM + MP) LSVM, 2009 [59]	13.20±0.62	15.45±0.37	16.37± 0.47	-	16.57±1.01
dense SIFT (KSRSPM) LSVM, 2010 [64]	29.77±0.14	35.67±0.10	38.61±0.19	-	40.30±0.22
dense SIFT (KC) SVM with HI, 2008 [12]	-	27.17 (25)*	-	-	-
dense SIFT (EMK) LSVM, 2009 [66]	23.2±0.6	30.5±0.4	34.4±0.4		37.6±0.5
dense SIFT (ScSPM + MP) LSVM, 2009 [59]	27.73±0.51	34.02±0.35	37.46±0.55	-	40.14±0.91
dense SIFT (LLC + MP) LSVM, 2010 [61]	34.36	41.19	45.31	-	47.68
dense SIFT (LSA + MMP) LSVM, 2011 [67]	-	-	-	-	36.52±0.26
dense SIFT (LDC + LLC/LSA + MP) LSVM, 2013 [63]	-	-	-	-	38.25±0.08
dense SIFT (LScSPM + MP) LSVM, 2013 [65]	29.99±0.15	35.74±0.10	38.47±0.51	-	40.32±0.32
dense SIFT (SSC + MP) OCL, 2012 [62]	30.59±0.35	37.08±0.36	40.68±0.16	-	43.48±0.38
CSIFT (LLC + MP) LSVM, 2015 [85]	28.58±0.35	35.20±0.36	38.97±0.16	-	41.31±0.38
SIFT + SVM, 2004 [30, 87]	-	-	-	29.4	-
HOG + SVM, 2005 [31, 87]	-	-	-	33.3	-
HOG (SPM) SVM, 2014 [31, 87]	-	-	-	32.7	-
LBP + SVM, 2002 [1, 87]	-	-	-	20.7	-
LBP (SPM) SVM, 2014 [1, 87]	-	-	-	20.5	-
dense SIFT (SPM) SVM, 2007 [11]	28.30	34.10	-	-	-
dense SIFT + NBNN, 2008 [58]	30.4 (25)*	36.0 (25)*	-	-	-
LGP (SPM) LSVM, 2013	22.86±0.41	28.89±0.33	31.13±0.28	32.02±0.29	33.14±0.51
OC-LBP (BoF) LSVM, 2013	25.77±0.22	31.28±0.26	34.91±0.27	35.52±0.29	37.83±0.32
LAID (SPM) LSVM, 2013	19.71±0.33	25.45±0.36	29.08±0.29	30.25±0.44	32.65±0.4
CLBP_S/M/C (SPM) LSVM, 2010	15.72±0.17	20.48±0.34	24.08±0.35	25.16±0.45	27.56±0.5
LTP (SPM) LSVM, 2010	23.12±0.26	29.33±0.27	31.74±0.35	32.95±0.37	33.97±0.43
GIST + LSVM, 2001	18.58±0.27	21.36±0.15	24.17±0.12	26.14±0.29	27.09±0.5
CENTRIST (SPM) LSVM, 2011	21±0.34	27.13±0.29	29.97±0.31	31.12±0.43	32.72±0.82
Proposed (DTCTH + LSVM)	27.43±0.37	33.57±0.43	36.38±0.33	37.59±0.35	38.30±0.31
Proposed (DTCTH + HI)	32.91±0.31	39.42±0.21	43.07±0.18	44.16±0.25	45.61±0.27

* Different number of test images used for the experiment rather than standard settings

and pose than that in Caltech-101. We have evaluated our algorithm in different settings such as considering 15, 30, 45, and 60 training images per class and using the rest as test data unless otherwise stated.

Table 4 presents the experimental results of DTCTH as well as existing state of the art methods in the literature on Caltech-256 dataset which shows that the proposed DTCTH performs better compared to other basic feature descriptors including GIST [10], CENTRIST [7], LTP [2] and LGP [16]. Besides, Borji et al. [87] perform a comparative evaluation of different existing techniques such as SIFT [87], HOG [87], HOG pyramid [87], LBP [87] and LBP pyramid [87] on this dataset, all of which produce inferior results compared to DTCTH. Moreover, DTCTH achieves more than 11% and 17% accuracy improvements over CENTRIST and GIST respectively by considering HI kernel.

Furthermore, DTCTH performs better than different sparse and soft-assignment coding based approaches including ScSPM [59], KSRSPM [64], LScSPM [65], EMK [66], LSA [67], SSC [62] and LDC [63] except LLC [61]. This LLC shows slightly better result (73.44) compared to DTCTH (72.26) using LSVM with the cost of high-level representation. It is noteworthy that such high-level representation is computationally expensive. In contrast, the proposed DTCTH achieves comparable accuracy with much lower computation. A recent state of the art low-level descriptor is Reversal Invariant Descriptor Enhancement (RIDE) [88] that improves the performance of basic SIFT using a high-level representation that uses improved fisher vector (IFV) [89]. This IFV helps to boost-up of the performance and achieves 60.25% accuracy.

5.2 Scene Classification

We have implemented DTCTH for both indoor and outdoor scene classification. For this purpose, three datasets such as MIT Indoor 67 [93] for indoor, OT scene [10] for outdoor and Scene 15 [5] for both indoor and outdoor scene classification are used. The description of these

Table 5: Scene classification rate (%) in MIT Indoor 67

Techniques	Accuracy
CNN-SVM, 2014 [90]	58.4
Places-CNN, 2014 [70]	68.24
ImageNet-CNN, 2014 [70]	56.79
Hybride-CNN, 2014 [70]	70.80
dense SIFT (LSA + MMP) LSVM, 2011 [67]	44.19
dense SIFT (LLC + MP) LSVM, 2010 [61]	43.78
dense SIFT (LDC + LLC/LSA + MP) LSVM, 2013 [63]	46.69
dense SIFT (SSC + MP) OCL, 2012 [62]	44.35
Object Bank + LSVM, 2010 [54]	37.60
dense SIFT (BoF) SVM with HI, 2014 [91]	45.86
DPM, 2011 [55]	30.40
CENTRIST (BoF) PmSVM-HI, 2012 [78]	47.15
CENTRIST (BoF) PmSVM-χ^2, 2012 [78]	46.20
PRICoLBP + SVM with χ^2, 2014 [8]	43.4
HOG, 2005 [55]	22.8
SPM, 2006 [8],	34.4
MM-scene, 2010 [92]	28.00
mCENTRIST (SPM) LSVM, 2014 [6]	44.6±1.2
mSIFT (SPM) LSVM, 2014 [6]	39.7±1.6
mGIST (SPM) LSVM, 2014 [6]	31.5±1.6
LGP (SPM) LSVM, 2013	34.24±1.12
OC-LBP (BoF) LSVM, 2013	36.99±2.34
LAID (SPM) LSVM, 2013	32.78±1.47
CLBP_S/M/C (SPM) LSVM, 2010	30.45±1.70
LTP (SPM) LSVM, 2010	35.87±1.23
GIST + LSVM, 2001	26.5±1.41
CENTRIST (SPM) LSVM, 2011	35.12±0.99
Proposed (DTCTH + LSVM)	43.33±0.72
Proposed (DTCTH + HI)	46.22±1.02

Table 6: Scene classification rate (%) in OT scene

Techniques	Accuracy
dense color SIFT (pLSA) KNN, 2006 [68]	86.65
dense color SIFT (pLSA) SVM, 2008 [34]	82.50
dense color SIFT (SP-pLSA) SVM, 2008 [34]	87.8
HSOG (LLC + MP) SVM, 2014 [9]	86.3 *
dense SIFT (BoF) SVM, 2004 [9, 30]	84.1 *
HOG (BoF) SVM, 2005 [9, 31]	82.4 *
DAISY (BoF) SVM, 2010 [9, 45]	85.7 *
$CS-LBP_{2,8,0.01}$ (BoF) SVM, 2009 [9, 86]	83.4 *
dense color SIFT (SPM) SVM, 2008 [34]	87.1
LGP (SPM) LSVM, 2013	84.52
OC-LBP (BoF) LSVM, 2013	84.67
LAID (SPM) LSVM, 2013	84.25
CLBP_S/M/C (SPM) LSVM, 2010	79.34
LTP (SPM) LSVM, 2010	85.6
GIST + LSVM, 2001	69.03
CENTRIST (SPM) LSVM, 2011	84.01
Proposed (DTCTH + LSVM)	87.88±0.51
Proposed (DTCTH + HI)	89.18±0.81

* Half of the images for training and another half for testing.

three datasets are discussed below followed by the experimental results.

MIT Indoor 67. This dataset holds 15,620 images of 67 indoor scene categories [93]. There are at least 100 images in each category. We randomly choose 80 images from each category for training and remaining images for testing the system.

OT Scene. Oliva and Torralba first used OT scene dataset for scene classification [10]. It consists of 2,688 images from 8 scene classes. In the experiments, 100 images are randomly selected to train the system and the other images are used for testing purpose.

Scene 15. Scene 15 dataset contains 4,485 images of 15 scene categories [5]. Each category has between 200 and 400 images. We randomly select 100 images from each category as training data and use the remaining images as test data.

In general, indoor scene classification is comparatively challenging than outdoor scene classification because indoor scenes contain large inter-class similarity. Therefore, the performance of all the methods are generally lower for indoor scene (e.g., MIT Indoor 67) compared to the outdoor scene (e.g., OT Scene) datasets. Several state of the art low-level feature descriptors such as PRICoLBP [8], CENTRIST [7], GIST [10], SIFT [30], HOG [31], HSOG [9], CS-LBP [86], LTP [2] and LGP [16] are used for both indoor and outdoor scene classification. Recently, CENTRIST has been extended to multiple channels (mCENTRIST) [6], which shows better result (44.6%) in scene classification than CENTRIST (35.12%). They have also showed that multi-channel GIST (mGIST) and multi-channel SIFT (mSIFT) perform better than original GIST and SIFT respectively. DTCTH obtains better accuracy than all of these approaches in all the datasets (see Table 5, Table 6 and Table 7).

Besides these basic features, there are other methods such as NBNN [58], PmSVM [78], pLSA [68], SP-pLSA [34], Bag-of-Phrase (BoP) [95] and DAISY [45] are also used for scene classification. To this end, DTCTH achieves better results in the respective datasets than most of these approaches. In few cases, such as SP-pLSA shows slightly better results (83.7%) for Scene 15 dataset compared to DTCTH (83.63%). However, DTCTH achieves higher accuracy (89.18%) compared to SP-pLSA (87.8%) in OT Scene dataset. BoP uses histogram mining with discriminating learning technique, and achieves 86.78% accuracy in Scene 15 dataset. RIDE achieves 64.93% accuracy on MIT Indoor 67 dataset by adopting IFV which is computationally expensive as described earlier [88]. In OT scene dataset, DTCTH achieves the highest correct classification rate (89.18%).

Table 7: Scene classification rate (%) in Scene 15

Techniques	Accuracy
Places-CNN [70]	90.19
ImageNet-CNN [70]	84.23
Hybride-CNN [70]	91.59
SIFT (SPM + pLSA) SVM, 2006 [5]	81.40±0.50
dense color SIFT (pLSA) SVM, 2008 [34]	72.7
dense color SIFT (SP-pLSA) SVM, 2008 [34]	83.7
dense SIFT (KC) SVM with HI, 2008 [12]	77.10
dense SIFT (LSPM + MP) LSVM, 2009 [59]	65.32±1.02
dense SIFT (ScSPM + MP) LSVM, 2009 [59]	80.28±0.93
dense SIFT (Sparse Code) LSVM, 2010 [94]	84.10±0.50
dense SIFT (LLC + MP) LSVM, 2010 [61, 63]	79.81±0.35
dense SIFT (LSA + MMP) LSVM, 2011 [67]	82.70±0.39
SIFT (BOVW + SPCK++) SVM, 2011 [56]	82.51±0.43
dense SIFT (LDC + LLC/LSA + MP) LSVM, 2013 [63]	82.50±0.47
dense SIFT (LCSR + MP) LSVM, 2012 [60]	82.67±0.51
Object Bank + LSVM, 2010 [54]	80.90
dense SIFT + I2CDML, 2010 [57]	77.00±0.6
dense SIFT (SPM) I2CDML, 2010 [57]	81.2±0.52
dense SIFT + NBNN, 2008 [57, 58]	72.3±0.93
PRICoLBP + SVM with χ^2, 2014 [8]	82.04
dense SIFT (BoF) SVM with HI, 2014 [91]	82.06
LGP (SPM) LSVM, 2013	78.22±0.56
OC-LBP (BoF) LSVM, 2013	77.22±0.4
LAID (SPM) LSVM, 2013	81.18±0.6
CLBP_S/M/C (SPM) LSVM, 2010	76.47±0.15
LTP (SPM) LSVM, 2010	80.25±0.31
GIST + LSVM, 2001	55.55±0.67
CENTRIST (SPM) LSVM, 2011	81.45±0.23
Proposed (DTCTH + LSVM)	82.66±0.5
Proposed (DTCTH + HI)	83.63±0.21

In this dataset, comparing with GIST which is designed for scene classification, DTCTH increases the performance over 18% and 20% by considering linear and HI kernel respectively. DTCTH provides 83.63% accuracy in Scene 15 dataset which also demonstrates 2% and 28% improvements over CENTRIST and GIST respectively. Furthermore, DTCTH outperforms Object Bank, DPM, SPCK++, and NBNN in the respective datasets.

Considering high-level image representation, sparse and soft-assignment coding based approaches are well-known. Among these approaches, ScSPM [59], LLC [61], SSC [62], LSA [67], LCSR [60] and LDC [63] have gained popularity for scene classification. Most of these approaches use two steps for feature representation such as feature encoding and pooling (e.g., average, max) steps. Boureau et al. [94] perform a comparative experimental analysis which shows that sparse coding with MP achieves better result than other combinations in Scene 15. Among all of these approaches, LDC [63] achieves slightly better classification accuracy (46.69%) than DTCTH (46.22%) in MIT indoor 67, but this method produces inferior results comparing with DTCTH in Caltech-101 (4.09% inferior), Caltech-256 (7.36% inferior) and Scene 15 (1.13% inferior) datasets.

Table 8: Event classification rate (%) in UIUC Sports Event

Techniques	Accuracy
Places-CNN, 2014 [70]	94.12
ImageNet-CNN, 2014 [70]	94.42
Hybride-CNN, 2014 [70]	94.22
dense SIFT (KSRSPM) LSVM, 2010 [64]	84.92±0.78
dense SIFT (ScSPM + MP) LSVM, 2009 [59]	82.74±1.46
dense SIFT (LSA + MMP) LSVM, 2011 [67]	82.29±1.84
dense SIFT (LLC + MP) LSVM, 2010 [61]	81.41±1.84
dense SIFT (LCSR + MP) LSVM, 2012 [60]	87.23±1.14
dense SIFT + I2CDML, 2010 [57]	78.5±1.63
dense SIFT (SPM) I2CDML, 2010 [57]	79.7±1.83
dense SIFT + NBNN, 2008 [57, 58]	67.6±1.1
dense SIFT (BoF) SVM with HI, 2014 [30, 91]	85.12
LQP + SVM with RBF, 2012 [39, 51]	78.9
DDLBP + Max Relevance + SVM with RBF, 2014 [39]	83.5
DDLBP + mRMR + SVM with RBF, 2014 [39]	83.5
DDLBP + MJMI + SVM with RBF, 2014 [39]	84.0
mGIST (SPM) LSVM, 2014 [6]	76.2±1.9
mSIFT (SPM) LSVM, 2014 [6]	84.2±0.7
mCENTRIST (SPM) LSVM, 2014 [6]	86.5±0.6
LGP (SPM) LSVM, 2013	78.42±0.94
OC-LBP (BoF) LSVM, 2013	81.15±2.18
LAID (SPM) LSVM, 2013	78.50±0.65
CLBP_S/M/C (SPM) LSVM, 2010	78.88±0.92
LTP (SPM) LSVM, 2010	82.43±1.17
GIST + LSVM, 2001	69.95±0.98
CENTRIST (SPM) LSVM, 2011	79.50±0.95
Proposed (DTCTH + LSVM)	85.16±0.96
Proposed (DTCTH + HI)	88.18±0.84

5.3 Event Classification

The description of the dataset followed by experimental results are discussed in the following.

UIUC Sports Event. This dataset consists of 1,579 images of 8 sports event categories [33]. The number of images in each class ranges from 137 to 250. We

Table 9: Leaf classification rate (%) in Swedish Leaf

Techniques	Accuracy	Input
Soderkvist, 2001 [96]	82.40	Contour only
SC + DP, 2007 [97]	88.12	Contour only
IDSC + DP, 2007 [97]	94.13	Contour only
SPTC + DP, 2007 [97]	95.33	Gray-scale
Shape-Tree, 2007 [98]	96.28	Contour only
CENTRIST, 2011 [7, 8]	90.61	Contour only
SLPA, 2013 [99]	96.33	Gray-scale
PRICoLBP + SVM with χ^2, 2014 [8]	99.38	Gray-scale
LTP (SPM) LSVM, 2010	98.20	Gray-scale
LGP (SPM) LSVM, 2013	98.08	Gray-scale
LAID (SPM) LSVM	99.33	Gray-scale
CLBP_S/M/C (SPM) LSVM	98.53	Gray-scale
OC-LBP (BoF) LSVM	99.36	Gray-scale
GIST + LSVM	96.08	Gray-scale
CENTRIST (SPM) LSVM	97.44	Gray-scale
Proposed (DTCTH + LSVM)	99.49	Gray-scale
Proposed (DTCTH + HI)	99.52	Gray-scale

have followed the experimental settings described in [76] which is, randomly selecting 70 images as the training and other 60 for testing.

DTCTH outperforms all the low-level descriptors (as described before) even mCENTRIST [6] that uses color information for this dataset (see Table 8). It also shows better results compared to many high-level representation (see Table 8) with few exceptions such as BoP (91.74%) that uses saliency map and mining strategy to boost-up its performance [95].

5.4 Leaf Classification

For leaf classification, we use Swedish Leaf dataset [96]. The dataset description followed by experimental results are discussed in the following.

Swedish Leaf. This dataset consists of 15 species of leaves with 75 images per species [96]. The dataset has two properties such as the leaf images are manually aligned well and in a good shape. Following the standard protocol discussed in [8], 25 randomly selected images form each species are used for training and the rest for testing.

Table 9 presents the experimental results of DTCTH as well as the existing techniques in literature on this dataset, which shows that DTCTH achieves 99.52% accuracy. Several techniques are used in this dataset for shape and leaf classification. DTCTH outperforms all of these approaches by considering gray-scale image as input which is provided in Table 9.

Table 10: Expression recognition rate (%) in CK

Techniques	CK	
	6-class expression	7-class expression
Ranzato et al. [100]	-	90.1
LBP, 2006 [13]	92.6 ± 2.9	88.9 ± 3.5
LBP + Template Matching, 2009 [18]	84.5 ± 5.2	79.1 ± 4.6
Geometric Feature + TAN, 2003 [101]	-	73.2
LBP + SVM, 2009 [18]	91.5±3.1	88.1±3.8
Boosted-LBP, 2009 [18]	89.8 ± 4.7	85.0 ±4.5
Boosted-LBP + SVM, 2009 [18]	95.0 ±3.2	91.1 ± 4.0
Gabor + SVM, 2003 [52]	-	84.8
Gabor, 2009 [18]	89.4 ± 3.0	86.6 ± 4.1
LDN + LSVM, 2013 [50]	98.4±1.4	92.3±3.0
LGP (SPM) LSVM, 2013	93.36±3.76	88.97±4.18
OC-LBP (BoF) LSVM, 2013	84.84±5.29	78.17±5.50
LAID (SPM) LSVM, 2013	89.13±5.41	84.21±4.73
CLBP_S/M/C (SPM) LSVM, 2010	85.44±4.92	78.59±5.78
LTP (SPM) LSVM, 2010	91.18±8.68	88.79±2.31
CENTRIST (SPM) LSVM, 2011	89.84±7.90	86.69±2.04
Proposed (DTCTH + LSVM)	98.98±1.29	92.75±5.43
Proposed (DTCTH + HI)	97.76±2.43	93.89±2.63

5.5 Facial Expression Recognition

We also evaluate the performance of DTCTH in expression recognition. Most of the facial expression recognition systems attempt to recognize a set of expressions like anger, disgust, fear, joy, sadness and surprise. This 6-class expression set can also be extended to a 7-class expression set including a neutral expression. In this work, our aim is to recognize both 6-class and 7-class expressions. For this purpose, we have performed experiments on Cohn Kanade (CK) [81] and CK+ [102] datasets, where person independent 10 folds cross-validation testing is considered. More specifically, the whole dataset is divided into ten person independent groups of roughly equal number of subjects. Nine groups are used to train the classifier, and the remaining group is used as the test data. The datasets description along with experimental results are discussed in the following.

CK and CK+ Dataset. The CK dataset consists of 100 university students who were between 18 to 30 years old at the time of their inclusion. Among them, 65% are female. In the experimental setup, 320 image sequences are selected from 96 subjects, each of which is labeled as one of the six basic expressions. For 6-class expression recognition, the three most expressive image frames are taken from each sequence that results in 960 expression images. In order to build the neutral expres-

Table 11: 7 class expression recognition rate (%) in CK+

Techniques	Accuracy
AUDN, 2013 [103]	92.05
SPTS, 2006 [68]	50.4
CAPP, 2006 [68]	66.7
SPTS + CAPP, 2006 [68]	83.3
LDN + LSVM, 2013 [50]	89.3
NABP + Adaboost, 2015 [17]	92.17
LBP + Adaboost, 2006 [17]	88.67
LTP + Adaboost, 2010 [17]	89.65
LGP + Adaboost, 2013 [17]	83.10
HOG + Adaboost, 2005 [17]	89.69
OC-LBP + BoF + LSVM, 2013	84.20±4.90
LAID (SPM) LSVM, 2013	92.76
CLBP_S/M/C (SPM) LSVM, 2010	87.47
CENTRIST (SPM) LSVM	88.70±4.37
Proposed (DTCTH + LSVM)	93.99±5.83
Proposed (DTCTH + HI)	93.82±5.52

Table 12: Confusion matrix of DTCTH in case of 6-class expression recognition on CK

	Anger	Disgust	Fear	Sadness	Happy	Surprise
Anger	**99.22**	0.0	0.78	0.0	0.0	0.0
Disgust	0.0	**100.0**	0.0	0.0	0.0	0.0
Fear	0.0	0.0	**97.22**	0.0	2.78	0.0
Sadness	0.83	0.0	0.0	**98.33**	0.0	0.83
Happy	0.43	0.0	0.85	0.0	**98.72**	0.0
Surprise	0.0	0.0	0.0	0.0	0.0	**100.0**

Table 13: Confusion matrix of DTCTH in case of 7-class expression recognition on CK

	Anger	Disgust	Fear	Sadness	Happy	Neutral	Surprise
Anger	**86.67**	0.0	1.90	1.9	0.0	9.52	0.0
Disgust	0.77	**95.38**	0.0	0.0	0.0	3.85	0.0
Fear	0.56	0.0	**95.0**	0.0	0.56	3.89	0.0
Sadness	1.67	0.0	0.0	**93.9**	0.0	3.89	0.56
Happy	0.42	0.0	0.0	0.0	**99.2**	0.42	0.0
Neutral	2.71	0.21	1.67	0.21	0.63	**94.58**	0.0
Surprise	0.0	0.0	0.0	0.0	0.0	0.0	**100**

sion set, the first frame (i.e., neutral expression) from all 320 sequences are selected to make the 7-class expression dataset (1,280 images). Furthermore, the extended CK (CK+) is used, which includes 593 sequences for seven basic expressions including happiness, sadness, surprise, anger, disgust, fear and contempt. In the experiments, we select the most expressive three image frames from 327 sequences of 118 subjects.

Table 14: Confusion matrix of DTCTH in case of 7-class expression recognition on CK+

	Anger	Contempt	Disgust	Fear	Sadness	Happy	Surprise
Anger	**96.3**	2.22	0.74	0.0	0.74	0.0	0.0
Contempt	9.26	**87.04**	0.0	0.0	0.0	0.0	0.0
Disgust	0.56	0.0	**99.44**	0.0	0.0	0.0	0.0
Fear	0.0	0.0	1.33	**90.67**	0.0	8.0	0.0
Sadness	11.9	1.19	0.0	0.0	**85.7**	0.0	1.19
Happy	0.0	0.0	0.0	0.97	0.0	**99.03**	0.0
Surprise	0.0	0.0	0.40	0.0	0.0	0.0	**99.6**

DTCTH achieves better performance on CK (98.98%) and CK+ (93.99%) datasets than LBP [13], boosted LBP [18], NABP [17], LGP [16], LTP [2], HOG [31], LDN [50] and CENTRIST [7] which are presented in Table 10 and Table 11. DTCTH also achieves better expression recognition performance with lower computational cost (see Table 10).

Table 12 demonstrates the confusion matrix of 6 different expressions in CK dataset. From this matrix, it can be seen that DTCTH performs better in all the basic expressions. Anger, sadness and fear show comparatively lower performance than other expressions which is generally happened in expression recognition in CK dataset (see Table 13). However, other existing approaches provide inferior results in these expressions than DTCTH.

Table 11 presents the results of DTCTH in CK+ dataset which shows that DTCTH outperforms existing state of the art approaches such as LDN, NABP, LTP, LBP, LGP, HOG and CENTRIST. It is noteworthy to mention here that, DTCTH outperforms even deep learning based methods described in [100] and [103] on both CK and CK+ datasets. Besides this, Table 14 demonstrates the confusion matrix of seven different expressions in CK+ dataset. From this matrix, it can be concluded that DTCTH achieves better accuracy in challenging expressions such as contempt, sadness, fear and anger, though most of the existing techniques provide poor performance in these expressions.

6 Conclusion and Future Work

In this paper, a low-level feature representation technique namely Discriminative Ternary Census Transform Histogram (DTCTH) is proposed where we have showed the requirements of a low-level descriptor and introduce a way to achieve those. Rigorous experiments on five different applications including nine different datasets demonstrate that DTCTH have more discriminating

ability than other existing state of the art low-level descriptors. It even outperforms the methods that include several high-level representations for different applications. This is because DTCTH has the ability to capture the prominent features that are stable in the presence of noise and different lightening conditions.

For calculating the threshold of DTCTH, we describe a way that combines Jenk's amd Weber's law. We also provide a low cost approximation that we have found empirically. Further research on this issue is also required to obtain a more close low cost approximation. Moreover, the incorporation of color information and high-level feature representation like sparse coding and pooling might further boost the performance of this descriptor which will be addressed in future.

Competing interests

The authors declare that they have no competing interests.

References

1. Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.
2. Xiaoyang Tan and Bill Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Image Processing, IEEE Transactions on*, 19(6):1635–1650, 2010.
3. Mohammad Shoyaib, M Abdullah-Al-Wadud, and Oksam Chae. A noise-aware coding scheme for texture classification. *Sensors*, 11(8):8028–8044, 2011.
4. Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *Computer Vision—ECCV'94*, pages 151–158. Springer, 1994.
5. Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.
6. Yang Xiao, Jianxin Wu, and Junsong Yuan. mcentrist: A multi-channel feature generation mechanism for scene categorization. *Image Processing, IEEE Transactions on*, 23(2):823–836, 2014.
7. Jianxin Wu and James M Rehg. Centrist: A visual descriptor for scene categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1489–1501, 2011.
8. Xianbiao Qi, Rong Xiao, Chun-Guang Li, Yu Qiao, Jun Guo, and Xiaoou Tang. Pairwise rotation invariant co-occurrence local binary pattern. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(11):2199–2213, 2014.
9. Di Huang, Chao Zhu, Yunhong Wang, and Liming Chen. Hsog: a novel local image descriptor based on histograms of the second-order gradients. *Image Processing, IEEE Transactions on*, 23(11):4680–4695, 2014.
10. Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.
11. Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.
12. Jan C van Gemert, Jan-Mark Geusebroek, Cor J Veenman, and Arnold WM Smeulders. Kernel codebooks for scene categorization. In *Computer Vision–ECCV 2008*, pages 696–709. Springer, 2008.
13. Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037–2041, 2006.
14. Daniel Maturana, Domingo Mery, and Alvaro Soto. Learning discriminative local binary patterns for face recognition. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 470–475. IEEE, 2011.
15. Lun Zhang, Rufeng Chu, Shiming Xiang, Shengcai Liao, and Stan Z Li. Face detection based on multi-block lbp representation. In *Advances in biometrics*, pages 11–18. Springer, 2007.
16. Bongjin Jun, Inho Choi, and Daijin Kim. Local transform features and hybridization for accurate face and human detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(6):1423–1436, 2013.
17. Md Mostafijur Rahman, Shanto Rahman, Minhas Kamal, Emon Kumar Dey, M. Abdullah-Al-Wadud, and Mohammad Shoyaib. Noise adaptive binary pattern for face image analysis. In *Computer and Information Technology (ICCIT), 2015 18th International Conference on*. IEEE, 2015.
18. Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on

local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.
19. Mohammad Shoyaib, Jo Moo Youl, Muhammad Mahbub Alam, Oksam Chae, et al. Facial expression recognition based on a weighted local binary pattern. In *2010 13th International Conference on Computer and Information Technology (ICCIT)*, pages 321–324, 2010.
20. Md Mostafijur Rahman, Shanto Rahman, Emon Kumar Dey, and Mohammad Shoyaib. A gender recognition approach with an embedded preprocessing. *International Journal of Information Technology and Computer Science (IJITCS)*, 7(7):19, 2015.
21. Subrahmanyam Murala, RP Maheshwari, and R Balasubramanian. Local tetra patterns: a new feature descriptor for content-based image retrieval. *Image Processing, IEEE Transactions on*, 21(5):2874–2886, 2012.
22. Robert T Collins, Alan Lipton, Takeo Kanade, Hironobu Fujiyoshi, David Duggins, Yanghai Tsin, David Tolliver, Nobuyoshi Enomoto, Osamu Hasegawa, Peter Burt, et al. *A system for video surveillance and monitoring*, volume 2. Carnegie Mellon University, the Robotics Institute Pittsburg, 2000.
23. Vladimir Pavlovic, Rajeev Sharma, Thomas S Huang, et al. Visual interpretation of hand gestures for human-computer interaction: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):677–695, 1997.
24. Aditya Vailaya, Mário AT Figueiredo, Anil K Jain, and Hong-Jiang Zhang. Image classification for content-based indexing. *Image Processing, IEEE Transactions on*, 10(1):117–130, 2001.
25. Anil K Jain, Arun Ross, and Salil Prabhakar. An introduction to biometric recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(1):4–20, 2004.
26. Atam P Dhawan. *Medical image analysis*, volume 31. John Wiley & Sons, 2011.
27. Shanto Rahman, Md Mostafijur Rahman, Khalid Hussain, Shah Mostafa Khaled, and Mohammad Shoyaib. Image enhancement in spatial domain: A comprehensive study. In *Computer and Information Technology (ICCIT), 2014 17th International Conference on*, pages 368–373. IEEE, 2014.
28. Khalid Hussain, Shanto Rahman, SM Khaled, M Abdullah-Al-Wadud, and Mohammad Shoyaib. Dark image enhancement by locally transformed histogram. In *Software, Knowledge, Information Management and Applications (SKIMA), 2014 8th International Conference on*, pages 1–7. IEEE, 2014.
29. Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. Otc: A novel local descriptor for scene classification. In *Computer Vision–ECCV 2014*, pages 377–391. Springer, 2014.
30. David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
31. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.
32. Oscar Déniz, Gloria Bueno, Jesús Salido, and Fernando De la Torre. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, 32(12):1598–1603, 2011.
33. Li-Jia Li and Li Fei-Fei. What, where and who? classifying events by scene and object recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.
34. Anna Bosch, Andrew Zisserman, and Xavier Muoz. Scene classification using a hybrid generative/discriminative approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(4):712–727, 2008.
35. Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.
36. Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630, 2005.
37. Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–506. IEEE, 2004.
38. Yadong Mu, Shuicheng Yan, Yi Liu, Thomas Huang, and Bingfeng Zhou. Discriminative local binary patterns for human detection in personal album. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
39. Jianfeng Ren, Xudong Jiang, Junsong Yuan, and Gang Wang. Optimizing lbp structure for visual recognition using binary quadratic programming. *Signal Processing Letters, IEEE*, 21(11):1346–1350, 2014.

40. Chao Zhu, Charles-Edmond Bichot, and Liming Chen. Image region description using orthogonal combination of local binary patterns enhanced with color information. *Pattern Recognition*, 46(7):1949–1963, 2013.
41. Di Huang, Caifeng Shan, Mohsen Ardabilian, Yunhong Wang, and Liming Chen. Local binary patterns and its application to facial image analysis: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):765–781, 2011.
42. SM Zahid Ishraque, Mohammad Shoyaib, Mohammad Abdullah-Al-Wadud, Md Monirul Hoque, and Oksam Chae. A local adaptive image descriptor. *New Review of Hypermedia and Multimedia*, 19(3-4):286–298, 2013.
43. Onkar Dabeer and Subhasis Chaudhuri. Analysis of an adaptive sampler based on weber's law. *IEEE Transactions on Signal Processing*, 59(4):1868–1878, 2011.
44. Zhenhua Guo, Lei Zhang, and David Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, 2010.
45. Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):815–830, 2010.
46. Matthew Brown, Gang Hua, and Simon Winder. Discriminative learning of local image descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):43–57, 2011.
47. Lois Robertson. Methods and innovations for multimedia database content management/current trends and future practices for digital literacy and competence. *The Australian Library Journal*, 62(2):170–171, 2013.
48. Yunqian Ma and Petar Cisar. Event detection using local binary pattern based dynamic textures. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 38–44. IEEE, 2009.
49. Mohammad Shoyaib, Mohammad Abdullah-Al-Wadud, SM Zahid Ishraque, and Oksam Chae. Facial expression classification based on dempster-shafer theory of evidence. In *Belief Functions: Theory and Applications*, pages 213–220. Springer, 2012.
50. Adin Ramirez Rivera, Jorge Rojas Castillo, and Oksam Chae. Local directional number pattern for face analysis: Face and expression recognition. *Image Processing, IEEE Transactions on*, 22(5):1740–1752, 2013.
51. Sibt ul Hussain and Bill Triggs. Visual recognition using local quantized patterns. In *Computer Vision–ECCV 2012*, pages 716–729. Springer, 2012.
52. Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, volume 5, pages 53–53. IEEE, 2003.
53. Wenfei Gu, Cheng Xiang, YV Venkatesh, Dong Huang, and Hai Lin. Facial expression recognition using radial encoding of local gabor features and classifier synthesis. *Pattern Recognition*, 45(1):80–91, 2012.
54. Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386, 2010.
55. Megha Pandey and Svetlana Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1307–1314. IEEE, 2011.
56. Yi Yang and Shawn Newsam. Spatial pyramid co-occurrence for image classification. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1465–1472. IEEE, 2011.
57. Zhengxiang Wang, Yiqun Hu, and Liang-Tien Chia. Image-to-class distance metric learning for image classification. In *Computer Vision–ECCV 2010*, pages 706–719. Springer, 2010.
58. Oren Boiman, Eli Shechtman, and Michal Irani. In defense of nearest-neighbor based image classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
59. Jianchao Yang, Kai Yu, Yihong Gong, and Tingwen Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1794–1801. IEEE, 2009.
60. Aymen Shabou and Hervé LeBorgne. Locality-constrained and spatially regularized coding for scene categorization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3618–3625. IEEE, 2012.

61. Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010.
62. Gabriel L Oliveira, Erickson R Nascimento, Antonio W Vieira, and Mario FM Campos. Sparse spatial coding: A novel approach for efficient and accurate object recognition. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 2592–2598. IEEE, 2012.
63. Zilei Wang, Jiashi Feng, Shuicheng Yan, and Hongsheng Xi. Linear distance coding for image classification. *Image Processing, IEEE Transactions on*, 22(2):537–548, 2013.
64. Shenghua Gao, Ivor Wai-Hung Tsang, and Liang-Tien Chia. Kernel sparse representation for image classification and face recognition. In *Computer Vision–ECCV 2010*, pages 1–14. Springer, 2010.
65. Shenghua Gao, Ivor Wai-Hung Tsang, and Liang-Tien Chia. Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):92–104, 2013.
66. Liefeng Bo and Cristian Sminchisescu. Efficient match kernel between sets of features for visual recognition. In *Advances in neural information processing systems*, pages 135–143, 2009.
67. Lingqiao Liu, Lei Wang, and Xinwang Liu. In defense of soft-assignment coding. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2486–2493. IEEE, 2011.
68. Anna Bosch, Andrew Zisserman, and Xavier Muñoz. Scene classification via plsa. In *Computer Vision–ECCV 2006*, pages 517–530. Springer, 2006.
69. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
70. Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.
71. Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.
72. Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.
73. Liu Yang, Rong Jin, Rahul Sukthankar, and Frederic Jurie. Unifying discriminative visual codebook generation with classifier training for object category recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
74. Subhransu Maji, Alexander C Berg, and Jitendra Malik. Classification using intersection kernel support vector machines is efficient. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
75. Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2012.
76. Hao Zhang, Alexander C Berg, Michael Maire, and Jitendra Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2126–2136. IEEE, 2006.
77. John C Platt, Nello Cristianini, and John Shawe-Taylor. Large margin dags for multiclass classification. In *nips*, volume 12, pages 547–553, 1999.
78. Jianxin Wu. Power mean svm for large scale visual classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2344–2351. IEEE, 2012.
79. University of Kansas. Department of Geography and GF Jenks. *Optimal data classification for choropleth maps*. 1977.
80. Robert G Cromley. A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data. *International Journal of Geographical Information Systems*, 10(4):405–424, 1996.
81. Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53. IEEE, 2000.
82. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.
83. Prateek Jain, Brian Kulis, and Kristen Grauman. Fast image search for learned metrics. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8.

IEEE, 2008.
84. Giuseppe Serra, Costantino Grana, Marco Manfredi, and Rita Cucchiara. Gold: Gaussians of local descriptors for image representation. *Computer Vision and Image Understanding*, 134:22–32, 2015.
85. Junzhou Chen, Qing Li, Qiang Peng, and Kin Hong Wong. Csift based locality-constrained linear coding for image classification. *Pattern Analysis and Applications*, 18(2):441–450, 2015.
86. Marko Heikkilä, Matti Pietikäinen, and Cordelia Schmid. Description of interest regions with local binary patterns. *Pattern recognition*, 42(3):425–436, 2009.
87. Ali Borji and Laurent Itti. Human vs. computer in scene and object recognition. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 113–120. IEEE, 2014.
88. Lingxi Xie, Jingdong Wang, Weiyao Lin, Bo Zhang, and Qi Tian. Ride: Reversal invariant descriptor enhancement. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 100–108, 2015.
89. Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer, 2010.
90. Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.
91. Zhen Zuo, Gang Wang, Bing Shuai, Lifan Zhao, Qingxiong Yang, and Xudong Jiang. Learning discriminative and shareable features for scene classification. In *Computer Vision–ECCV 2014*, pages 552–568. Springer, 2014.
92. Jun Zhu, Li-Jia Li, Li Fei-Fei, and Eric P Xing. Large margin learning of upstream scene understanding models. In *Advances in Neural Information Processing Systems*, pages 2586–2594, 2010.
93. Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE, 2009.
94. Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce. Learning mid-level features for recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2559–2566. IEEE, 2010.
95. Baiying Lei, Ee-Leng Tan, Siping Chen, Dong Ni, and Tianfu Wang. Saliency-driven image classification method based on histogram mining and image score. *Pattern Recognition*, 48(8):2567–2580, 2015.
96. Oskar Söderkvist. Computer vision classification of leaves from swedish trees. 2001.
97. Haibin Ling and David W Jacobs. Shape classification using the inner-distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(2):286–299, 2007.
98. Pedro F Felzenszwalb and Joshua D Schwartz. Hierarchical matching of deformable shapes. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.
99. Shanwen Zhang, Yingke Lei, Tianbao Dong, and Xiao-Ping Zhang. Label propagation based supervised locality projection analysis for plant leaf classification. *Pattern Recognition*, 46(7):1891–1897, 2013.
100. Joshua Susskind, Volodymyr Mnih, Geoffrey Hinton, et al. On deep generative models with applications to recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2857–2864. IEEE, 2011.
101. Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, 91(1):160–187, 2003.
102. Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.
103. Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-aware deep networks for facial expression recognition. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.