


An effective fusion model for image retrieval

Leila Mansourian¹  · Muhamad Taufik Abdullah¹ ·
Lili Nurliyana Abdullah¹ · Azreen Azman¹ ·
Mas Rina Mustaffa¹

Received: 7 December 2015 / Revised: 19 July 2017 / Accepted: 4 September 2017 /
Published online: 18 October 2017
© Springer Science+Business Media, LLC 2017

Abstract In the past decade, the popular Bag of Visual Words approach has been applied to many computer vision tasks, including image classification, video search, robot localization, and texture recognition. Unfortunately, most approaches use intensity features and discard color information, an important characteristic of any image that is motivated by human vision. Besides, if background colors are higher than foreground ones, Dominant Color Descriptor (DCD) retrieves images that contain similar background colors correctly. On the other hand, just color feature extraction is not sufficient for similar objects with different color descriptors (e.g. white dog vs. black dog). To solve these problems, a new Salient DCD (SDCD) color descriptor is proposed to extract foreground color and add semantic information into DCD based on the color distances and salient object extraction methods. Besides, a new fusion model is presented to fuse SDCD histogram and PHOW MSDSIFT histogram. Performance evaluation on several datasets proves that the new approach outperforms other existing, state-of-the-art methods.

This article was kindly supported by the Malaysian Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS).

✉ Leila Mansourian
l.mansourian@yahoo.com

Muhamad Taufik Abdullah
mta@upm.edu.my

Lili Nurliyana Abdullah
liyana@upm.edu.my

Azreen Azman
azreenazman@upm.edu.my

Mas Rina Mustaffa
masrina@upm.edu.my

¹ Faculty of Computer Science and Information Technology, Department of Multimedia, University Putra Malaysia UPM, Serdang, Selangor Darul Ehsan, 43400, Malaysia

Keywords Saliency map · PHOW MSDSIFT feature · Bag of visual words model (BoVW) · Dominant color description (DCD) · Image retrieval · Pyramidal histogram of visual words (PHOW)

1 Introduction

There has been much recent interest, in both research and practice, in classifying images into categories. To achieve this goal, the first stage is keypoint extraction. Keypoints are salient image patches that contain rich local information of an image. There are different keypoint detectors, which are surveyed by Mikolajczyk and Schmid [43] and Zhang et al. [65]. Keypoints are depicted by descriptors such as the Scale-Invariant Feature Transform (SIFT).

Lowe [37] proposed SIFT, which is a robust feature in scaling, rotation, translation, and illumination, and is partially invariant to affine distortion. In addition, there is no need to digest images. The only thing we need to do is to quantize SIFT features by the well-known Bag of Visual Words (BoVW) technique, first presented by Csurka et al. [13].

The Bag of Words (BoW) model is a popular technique for document classification. In this method, a document is represented as a bag of its words, and features are extracted from the frequency of occurrence of each word. The Bag of Words model has also been used for computer vision by Perona [48]. Therefore, instead of document version name (BoW), Bag of Visual Words (BoVW) was used in the present research. For BoVW extraction, we must first extract blobs and features (e.g., SIFT). In the next stage, a visual vocabulary must be built by using a classification method (e.g., K-means). Representation of images with BoVW histograms is the third stage. The final stage is image classification, using a classification method (e.g., Support Vector Machine [SVM]).

O'Hara and Draper [46] presented a survey on BoVW image representations. They highlighted recent techniques that mitigated quantization errors, improved feature detection, and sped up image retrieval. Lazebnik et al. [30] presented an extension to the BoVW model for recognizing scene categories based on global geometric correspondence (spatial pyramid framework). Their method divides each image into sub-regions and computes the histograms of local features for each sub-region. This spatial pyramid was applied in later generation BoVW models, such as Ionescu et al. [20], which extracted dense SIFT descriptors of whole images or a spatial pyramid of the image. They also proposed a method for classifying human facial expressions from low-resolution images based on a bag of words representation. Pyramidal Histogram of Visual Words (PHOW) was proposed by Bosch et al. [10]. PHOW is an image descriptor based on SIFT feature. It uses a grid of dense points in the image, and a SIFT is computed for each point of the grid. By default, it uses three scales and builds a pyramid of descriptors.

Most of the above-mentioned models concentrate on grayscale versions of pictures and ignore the color information in pictures. Therefore, researchers have attempted to combine color features with other features to get better results. Vigo et al. [54] found that integrating color significantly improves the overall performance of both feature detection and extraction. Khan et al. [25] presented a method for object recognition by using multiple cues (shape and color). Their combination was based on modified shape features by category-specific color attention. Alqasrawi et al. [2] used a keypoint-density-based weighting method to combine a BoVW model with color information on a spatial pyramid layout. Recently, Barata et al. [4] compared grayscale methods against color sampling methods (Harris Laplacian detector and SIFT descriptor). They found that color detectors

and Color-SIFT perform better. Jalali et al. [22] utilized color to enhance object and scene recognition in a method inspired by the characteristics of color and object-selective neurons. A comprehensive discussion of the combination of color with Bag of Visual Words image representations may be found in Weijer et al. [58].

Some researchers have investigated the application of BoVW classification in special domains. For instance, the authors of this article investigated the potential use of the Bag of SIFT feature for animal classification and determined which classification method was better for animal pictures [40]. Ionescu et al. [20] proposed a method for classifying human facial expressions from low-resolution images based on a BoVW representation. Abdelkhalak and Zouaki [1] suggested a new descriptor for bird searches in images. They concatenated shape, first color moment (mean), and second color moments (variance), an early fusion of color and shape, to build a BoVW.

As it can be seen, color features are one of the future methods for concentration, and they are still being improved. In addition, color is one of the important characteristics of human vision. However, in the traditional version of DCD, if background colors are higher than foreground ones, images with similar background colors are retrieved wrongly as belonging to the same category. Also, a lonely color feature is not sufficient for the similar objects with different color information. Therefore, this paper presented a new Salient DCD to add semantic information, and reduce the background effect. Also, a new fusion model for fusing SDCD and PHOW MSDSIFT histogram is proposed.

Our SDCD & PHOW MSDSIFT model approach consists of six main steps. The first one is saliency map computation based on Jiang et al. [24], which discriminates the background from the main object. The second step is divided in two parallel stages: SDCD color extraction of the salient part and PHOW feature extraction of salient and original part. In the next step, their codebook is constructed in parallel by K-means classification. Again in parallel during the fourth stage, spatial histogram descriptors are quantized based on a binary tree in which every node is a k-dimensional point (KD-trees) to identify the visual words. A homogeneous kernel maps of the histograms is extracted. Finally, these histogram kernel maps are fused together with a new fusion model which is described in section 3 to obtain a superior visual word constructed from SDCD and PHOW features. To test our model, spatial histograms of visual words of test pictures were compared with spatial histograms of visual words based on SVM Chi square ($SVM\ CHI^2$). Because, $SVM\ CHI^2$ [42] shown better results in the literature and current researchers early experiments [40]. Subsequently, the appropriate concept names were extracted for the test images by assigning class labels for the test images.

The rest of the current paper is structured as follows. In Section 2, some materials and methods related to our research are reviewed. Section 3 introduces the SDCD algorithm and SDCD & PHOW fusion model for image retrieval. Section 4 presents the experimental setup. A discussion of the proposed model, research results, and usefulness of SDCD & PHOW fusion model are explained in Section 5. The paper concludes with some comments on future research in Section 6.

2 Materials and methods

Finding appropriate methods for image classification and feature extraction based on location is a recent and controversial endeavor [3, 27–29, 36, 45, 66]. In the traditional BoVW model, visual words are collected and treated in the same way, even though they may be from an important part or the background of a picture. This means that the classifier often

relies on visual words that fall in the background and merely describe the context of the object [47]. Also, background features have higher percentage than foreground ones, previous image classification methods did not add semantic location information to the features. They can retrieve images that contain similar background and not images with similar foreground. This means, that they are dependent on background feature which is not a useful information. On the other hand, color is not sufficient for similar objects with various colors such as white dog or black dog. Based on these problems, a SDCD algorithm to extract important colors of salient parts of pictures and a new SDCD & PHOW fusion model for fusing SDCD color features and PHOW MSDSIFT features are proposed. This model can collect visual words of the whole and salient parts of a picture. In what follows, we first briefly review common stages and materials for color extraction and image retrieval techniques.

2.1 Image segmentation

The first stage, but not a mandatory one, is image segmentation. The segmentation algorithm divides images into different parts based on feature similarity. Different segmentation approaches proposed in the literature are based on: background removing based, clustering based, grid based, model based, contour based, graph based, region growing based and salient based method. For a comprehensive segmentation review, readers are referred to [16]. In this study, the focus is on salient-based methods. Because of the object location, removing the background parts is an important stage. Recently, much research has designed various models to compute the saliency maps. There are five major research areas for detecting saliency in images: *Salient Object Detection Methods*, *Localization Salient Models*, *Aggregation and Optimization Salient Models*, *Active Salient Models*, and *Segmentation Salient Models*. These research areas are described in detail in the following paragraphs.

2.1.1 Salient object detection methods

Based on the survey research conducted by Borji et al., there are two attributes for detecting salient or interesting objects in images: Block-based vs. region based analysis and intrinsic cues vs. extrinsic cues [9].

- *Block-based vs. Region-based analysis:*
Block (i.e. pixels and patches) based is an early method of finding a salient object, while regions are a widespread generation with the development of superpixel algorithms.
- *Intrinsic cues vs. Extrinsic cues:*
The key difference is for using attributes from one image (i.e. Intrinsic cues) or similar cooperation images (e.g. user annotations, depth map, or statistical information) to facilitate detecting salient objects in the image (i.e. Extrinsic cues).

Based on the literature reviews and mentioned attributes most of the existing salient object detection approaches can be divided into three major categories, *block-based models with intrinsic cues*, *region-based models with intrinsic cues*, and *models with extrinsic cues*.

1. Block-based Models with Intrinsic Cues:

These models detect salient objects based on blocks (i.e. pixels or patches) with only utilizing intrinsic cues. Their drawbacks are: they detect high contrast edges as a salient object instead of the real salient object, and if the size of blocks is large, the boundary of the salient object is not protected very well. To control these problems successfully, new researchers considered more on region based maps. Because the number of regions is much less than the number of blocks better features can be extracted from regions.

2. Region-based Models with Intrinsic Cues:

In these models, the first input image is segmented into regions aligned with intensity edges and then regional saliency map computed. Three types of region extraction methods are used for saliency computation (Graph-based segmentation algorithm, mean-shift algorithm, or clustering quantization). The first advantage of this method in comparison with block-based is that for improving these models, there are several choices like backgroundness, objectness, focusness and boundary connectivity. Besides, regions give more advanced cues (e.g. color histogram). Another advantage of using regions instead of blocks (i.e. pixels or patches) is for computational cost because each image has far fewer regions than pixels, computation of regional saliency would be less than producing full-resolution saliency maps. Despite these advantages, the new generation will be using extrinsic cues. Jiang et al. proposed an approach based on multi-scale local region contrast, which calculates saliency values across multiple segmentations and combines these regional saliency values to get a pixel-wise saliency map [23].

3. Models with Extrinsic Cues:

These models help salient object extraction in images and videos. These cues can be derived from the ground truth annotations of the training images, similar images, the video sequence, a set of input images containing the common salient objects, depth maps, or light field images. Borji et al. [9] concluded that the DRFI, which is presented by Jiang et al. [24], is an extrinsic cue model. This model had been only trained on a small subset of MSRA5K, and it still consistently outperforms other methods on all datasets. Previous categorizations (Block-based vs. Region-based analysis and intrinsic cues vs. Extrinsic cues) were based on salient object detection.

2.1.2 Localization salient models

Borji et al. [9] mentioned that there exist some other researches whose main research effort is not about the saliency map computation; nor can it segment or localize salient objects directly with bounding boxes. They classified them as *Localization models*, *Segmentation Models*, *Aggregation and Optimization Models*, and *Active Models*. The output of these models is rectangles around the salient objects by converting the binary segmentations to bounding boxes. The most common approach is using a sliding window and classifying each of them as either a target or a background. For example, Lampert et al. [29] proposed an object localization method based on maximization of sub-images with branch and bound scheme, but their research cannot find two or more important objects in one picture. Another problem with using sliding windows occurred when the local image information is insufficient e.g. when the target is very small or highly occluded. In these cases, other parts of the picture will help us to classify the picture [45]. Therefore, K. Murphy et al. presented a combination model of local and global (gist) features of the scene. This would be useful for solving the previous problem. They found that local features alone would cause a lot of false positives. Sometimes the scale estimation is incorrect as well. Also, they concluded that using global features can correct the estimation and decrease the ambiguity caused by only using local object detection methods. However, the basic idea of previous approaches that at least one salient object exists in the input image may not always behold as some background images that contain no salient objects at all. Wang et al. [57], investigated the problem of detecting the existence and the place of salient objects on thumbnail images using random forest learning approach. Recently, current researchers proposed a Salient Based Bag of Visual Word model (SBBovW) to recognize difficult objects that have had

low accuracy in previous methods [41]. This method integrates SIFT features of the original and salient parts of pictures and fuses them together to generate better codebooks using bag of visual word method. Also, it can find object place based on the salient map automatically. However, it did not use any color information.

2.1.3 Aggregation and optimization salient models

These models try to combine some saliency maps and in order to form an accurate map to help the detection of salient objects. Borji et al. [8] proposed a standard saliency aggregation. Recently, Yan et al. [60] combined saliency maps based on the hierarchical segmentation to get a tree-structure graphical model from three layers of different scales. In this model, each node is related to a region. They concluded that hierarchical algorithms could select optimal weights for each region instead of global weighting superpixels.

2.1.4 Active salient models

These models combine two stages into one (the most salient object detection and segmentation). Recently, Borji [7] presented an active model, which can locate the salient object by finding the peak pixels of the fixation map. Then it segments the picture by superpixels. Their method can connect fixation prediction and salient object segmentation. Based on Mikolajczyk et al.'s [44] research on the different scale and affine invariant interest point detectors, the best results are obtained by the Hessian-Laplace and Salient regions method.

2.1.5 Segmentation salient models

In these models separating the salient object from the background is the main approach. Kim et al. [26] proposed a region detection approach. Their method used dense local region detectors to extract suitable features for object recognition and image matching. Having applied boundary-preserving local regions (BPLRs), they asserted that their method can find the connectivity of pixels, and it can save the object boundaries for foreground discovery and object classification. Wang et al. [19] presented a framework to segment the salient object by contextual cues usage automatically. Their method incorporates texture, luminance and color cues. Also, it measures the similarity between neighboring pixels and computes the edge probability map to label them as background/foreground. Recently, Jiang et al. [24] presented saliency estimation as a regression problem, and their method still consistently outperforms other saliency methods on all datasets. Therefore, we selected their method to generate a salient map.

2.2 Feature extraction

The next stage for image retrieval is feature extraction. In the following sections, different SIFT e.g., Speeded Up Robust Features (SURF), PHOW, and Pyramid Histogram of Oriented Gradient (PHOG) and color features are described.

2.3 SIFT and SURF

SIFT was first proposed by Lowe [37]. This feature has four parameters: keypoint center (x and y coordinates), scale (the radius of the region), and orientation (an angle expressed in radians). SIFT detector is invariant and robust to translation, rotations, and scaling, and

is partially invariant to affine distortion and illumination changes. Later, Bay et al. [5] proposed Speeded Up Robust Features (SURF), which is a quicker SIFT. Liu et al. [33] suggested a fast algorithm for the computation of a dense set of SIFT descriptors. Dalal et al. [14] used the Histogram of Oriented Gradient (HOG) descriptor for pedestrian detection. Pyramid HOG (PHOG) and PHOW, the new generation of SIFT features, are described below; for more information, the authors refer the reader to [10].

2.4 PHOW and PHOG

PHOW is a new trend of SIFT features proposed by Bosch et al. [10]. It uses dense SIFT under different scales and builds a pyramid of descriptors. PHOG is the edge version of PHOW, which means it gathers features of the edge-detected picture (e.g., Canny). The stages of PHOW and PHOG are depicted in Fig. 1. In forming the pyramid, the grid at level 1 has 2^l cells along each dimension. Consequently, level 0 is represented by an N -vector corresponding to the N bins of the histogram, level 1 by a $4N$ -vector, etc. The pyramid descriptor of the entire image (PHOW, PHOG) is a vector with dimensionality $N \sum_{l=0}^L 4^l$.

Therefore, we selected this type of SIFT instead of pure SIFT, which is not recommended by recent studies.

2.5 Color features

A color feature is another important feature that helps us recognize objects as a human does. The first step in color feature extraction is color space selection. There are several color spaces e.g., Red, Green, Blue (RGB), Cyan, Magenta, Yellow, and Black (CMYK), Hue, Saturation, Value (HSV) and an Adams chromatic valence color space which is proposed by the International Commission on Illumination (CIE) in 1976 (CIE Luv). Digital images are usually stored in RGB color space. Unfortunately, the color distance in RGB space does not represent perceptual color distance [62] (e.g., two colors with larger distance can be perceptually more similar than another two colors with smaller distance). Considering this drawback, CIE Luv space was selected for the present research, because it is a uniform color space in terms of color distance. MPEG-7 is a color descriptor proposed by Yamada [59]. It includes seven color descriptors (dominant colors, scalable color histogram, color structure, color layout, and a group of frames/picture (GoF/GoP) color). In MPEG-7, Dominant Color Descriptor (DCD) describes color distribution in an image or a region. This paper focuses on using DCD color based on the advantages of this color descriptor in reviewed paper of Zhang et al. [63]. Other color descriptors, such as color coherence vector (CCV), color correlogram, and color structure descriptor (CSD) are useful for whole image representation.

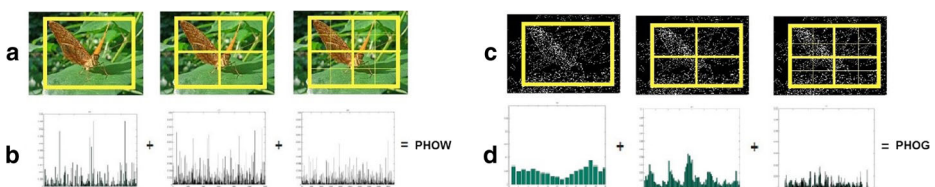


Fig. 1 Spatial SIFT representation (a,c) Grids for levels $l = 0$ to $l = 2$ for appearance and shape representation; (b,d) Appearance and SIFT histogram representations corresponding to each level

The DCD feature descriptor has two main components: (1) representative colors and (2) a percentage for each color. This descriptor is defined as:

$$F = \{c_i, p_i\}, i = 1, \dots, N \quad (1)$$

where N is the overall number of dominant colors for an image, c_i is a dominant color vector, p_i is the percentage for each dominant color, and the sum of p_i is equal to 1. MPEG-7 recommends the number of colors in a region and suggests that the value of N be in the range of 1 to 8.

Yang et al. [61] presented a fast, MPEG-7 dominant color extraction with a new similarity measure for image retrieval. In comparison to the previous versions of DCDs, it has higher accuracy and performance. According to Yang et al. [61], the distance between two images F_1 and F_2 is calculated by:

$$D^2(F_1, F_2) = 1 - SIM(F_1, F_2) \quad (2)$$

where $SIM(F_1, F_2)$ is the similarity measurement. This similarity measurement for two color features $F_1 = \{c_i, p_i\}, i = 1, \dots, N_1$ and $F_2 = \{c_i, p_i\}, i = 1, \dots, N_2$ is described as:

$$SIM(F_1, F_2) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} a_{ij} S_{ij} \quad (3)$$

where a_{ij} , is the coefficient of color similarity:

$$a_{ij} = \begin{cases} \frac{(1-d_{i,j})}{d_{max}} & d_{i,j} \leq T_d \\ 0 & d_{i,j} > T_d \end{cases} \quad (4)$$

where $d_{i,j}$, is the Euclidean distance between two color clusters c_i and b_j . Also, based on the research of Islam et al. [21], in CIE LUV color space, the value of T_d is fixed at 20. Because the dominant colors should be significant enough, we merge insignificant colors into nearby colors.

$$d_{(i,j)} = \|c_i - b_j\| \quad \text{and} \quad d_{max} = \alpha T_d \quad (5)$$

To properly reflect the similarity coefficient between two color clusters, the parameter α was set to 1 and $T_d = 20$ in the present research.

S_{ij} is a similarity score between two different dominant colors, given by:

$$S_{ij} = [1 - |p_q(i) - p_t(j)|] \times \min(p_q(i), p_t(j)) \quad (6)$$

where $p_q(i)$ and $p_t(j)$ are percentages of the i th and j th dominant color in query image and target image, respectively.

For dominant color vector quantization, an algorithm for automatic categorization was presented by Islam et al. [21]. They found that only 1.3% of image regions need more than 4 colors. For this reason, they shrink the number of dominant colors to four. With their algorithm, some of the salient regions cannot be properly described. In addition, their algorithm is very complicated, hard to implement, and very time-consuming. In 2013, [50] proposed WDCD for getting weight to each dominant color based on the salient map extraction. However, WDCD has a significant drawback for not retrieving similar objects with different color such as white dog and black dog. This means that color alone is not sufficient for image retrieval. To counteract these disadvantages, this paper presents a new, semantic base.

Which is a fast, and easy-to-understand dominant color vector quantization algorithm, and can find an appropriate number of colors based on their distances for extracting the DCD color of the salient part of a picture. This algorithm is described in Section 3. Later this color descriptor fused with another feature in a new fusion model to extract same objects with different colors.

2.6 Learning

In terms of training techniques, learning methods are divided into *Supervised*, *Unsupervised*, and *Semi-supervised (Hybrid)* models. They are the combination of clustering and classification techniques which are quickly growing [51].

In terms of feature selection, learning methods are divided into *Single view*, and *Multiview* feature extraction. These methods are described in the following paragraphs.

2.6.1 Single view learning

Single feature selection techniques are traditional learning methods that usually select features from a single task [39]. These methods have a basic drawback: They cannot precisely distinguish images containing several semantic concepts. Therefore, multiple feature selection methods are the new generations for feature extraction to eliminate the problem of single view feature extraction methods [39].

2.6.2 Multi view learning

Feature selection and feature transformation are two main ideas for feature extraction, but the former is the preferred method [34]. Although traditional feature selection methods prefer to use a single task, recent methods focus on multiple feature selection methods [34, 35, 38, 39]. Multi-task feature extraction handles correlated and noisy features. Even though all the features can be joined into a large vector, this strategy is not suitable and ignore the verity between features and may lead to severe cases of dimensionality [34].

A supervised multi-label multi-task feature learning is proposed by Wang et al. [55], but it is not suitable for classification.

In 2013, Liu and Tao [34] proposed a multiview Hessian Regularization (mHR) method for image annotation. Their method combines multiview features and Hessian regularization from different views.

Sparse coding finds a sparse linear combination of dictionary. It shows promising results for image denoising. In recent years, several of sparse coding algorithms have been developed [35]. The most noticeable sparse coding methods are based on Laplacian Regularization (LR). But LR based methods suffer from poor generality and only deal with a single view even though most of the times images are represented by multiple visual features [35]. Therefore, to overcome these drawbacks, W. Liu et al. applied multiview Hessian Discriminative Sparse Coding (mHDSC) to linear SVM and LS regression for image annotation. However, their method is tested on a small number of concepts (PASCAL VOC07 with 20 concepts).

In 2015, Y. Luo et al. proposed a multimodal multi-task feature extraction frame work (LM3FE) for classification which is suitable for image classification [38]. LM3FE uses all kinds of features even noisy features besides the complementarity of different modalities to reduce the redundancy in each modality. But their method is tested on small datasets (NUS-WIDE 12 concepts, and MIR 38 concepts).

Luo et al. [39] proposed a weight-based Matrix combining framework for transductive multilabel image classification. However the overall performance was not always satisfactory.

3 SDCD algorithm and SDCD & PHOW fusion model

In this section, the proposed algorithm and model are described in detail.

3.1 Proposed salient dominant color descriptor (SDCD) algorithm

Dominant Color Descriptor is an MPEG'7 color descriptor. DCD extracts colors of a region and based on the research of Zhang et al. [63], this color descriptor is useful for region color extraction and other color descriptors such as CCV, color correlogram, and CSD are useful for whole image representation.

However, in the traditional version of DCD, if background colors are higher than foreground ones, the algorithm retrieves images with similar background colors. Besides, color is not sufficient for similar objects with various colors (e.g. white dog, vs. black dog). Another drawback in the previous version of DCD, is that the maximum number of Dominant Colors was fixed at four [64]. Traditional segmentation methods (e.g. JSEG) create a lot of regions for each picture, and can not distinguish the most significant region or the important foreground region of the picture. WDCD which is proposed by [50] can not retrieve similar object with different colors (e.g. white dog or black dog).

To solve these problems in the previous DCD color descriptor, the salient region is extracted to extract colors of foreground or the important region of the picture. Therefore, the proposed algorithm wants to extract DCD colors of the salient part of pictures, so we may need more colors (more than four colors). For this reason, in this new algorithm, a Salient DCD for each region can have several colors based on the distances of the colors. SDCD combines semantic location information with DCD and removes background color which is not useful information. Moreover, implementation and understanding of this algorithm is easier. Later a fusion model is proposed to fuse SDCD color and PHOW MSDSIFT shape feature in order to retrieve similar objects with different colors accurately. SDCD algorithm is depicted in Algorithm 1. Its steps are:

1. Extract saliency map, based on Jiang et al. [24].
2. Extract salient region mask.
3. Clean the salient region mask from small spots.
4. If the mask is empty, this means that no salient region was found. Therefore, another mask with the size of the whole image will be created.
5. Multiply the mask with the original picture, and create a masked picture.
6. Find the line borders of the masked picture, and crop them.
7. Perform image smoothing and impulse noise removal with
8. peer group filtering (PGF) [15]. This algorithm swaps each image pixel with the weighted average of its peer group members, which are classified according to their color resemblance of neighboring pixels.
9. Classify colors of the smoothed picture into N colors.
10. Calculate the histogram of N colors, and divide them by summary to have a percentage of each color.
11. Calculate the Euclidean distances of the colors.

12. Merge near-color clusters (their distance is less than d_{max}).
13. Remove colors with small percentages of occurrence (less than 10 percent).

Algorithm 1 Proposed SDCD algorithm

```

1: procedure SDCD(Img)
2:   SalImg  $\leftarrow$  SalExt(Img);
3:   SalMsk  $\leftarrow$  SalMskExt(SalImg);
4:   SalMsk  $\leftarrow$  ClnSalMsk(SalMsk);
5:   if SalMsk =  $\emptyset$  then
6:     SalMsk  $\leftarrow$  true(size(Img));
7:   end if
8:   MskImg  $\leftarrow$  SalMsk  $\times$  Img;
9:   MskImg  $\leftarrow$  brdCrop(MskImg);
10:  NewImg  $\leftarrow$  imgPGF(MskImg);
11:  [CluIdx, CluCtr]  $\leftarrow$  kmeans(NewImg, N);
12:  [Cnt, Ctr]  $\leftarrow$  hist(CluIdx, N);
13:  PerCnt  $\leftarrow$  Cnt/sum(Cnt);
14:  CluDist  $\leftarrow$  eDist(cluCtr);
15:  [NewCluCtr, NewCluIdx, NewPerCnt, NewCnt]  $\leftarrow$ 
    mrgClu(CluDist, CluCtr, CluIdx, PerCnt, NewImg, Cnt);
16:  [NewCluCtr, NewCluIdx, NewPerCnt, NewCnt]  $\leftarrow$ 
    remSmallPercent(NewCluCnt, NewCluIdx, NewPerCnt, NewCnt)
17: end procedure

```

3.2 Proposed salient based fusion model (SDCD& PHOW MSDSIFT BoVW)

In the traditional BoVW model, a classifier often relies on visual words that fall in the background and merely describe the context of the object [47]. As mentioned before, [50] proposed WDCD which can not retrieve similar objects with different color (e.g. white dog or black dog). Based on this problem, the new SDCD & PHOW model for fusing SDCD color features and PHOW MSDSIFT features is proposed. This model can collect visual words from the whole and salient parts of a picture. The stages of the model are:

1. Saliency map and salient region mask computation
2. SDCD BoVW stages:
 - (a) Generate masked pictures
 - (b) Extract SDCD of the masked pictures
 - (c) Create SDCD codebook, based on K-Means classification
 - (d) Quantize SDCD visual words spatial histograms, based on KD-trees
 - (e) Transform non-linear histograms into a compact linear representation by Homogeneous Kernel Map.
3. PHOW MSDSIFT BoVW stages:
 - (a) Generate saliency rectangular parts of the picture
 - (b) Extract PHOW MSDSIFT feature from salient rectangular parts and normal pictures
 - (c) Create PHOW MSDSIFT codebook, based on K-Means classification

- (d) Quantize visual words spatial histograms, based on KD-trees
 - (e) Transform non-linear histograms into a compact linear representation by Homogeneous Kernel Map.
4. Histogram fusion by combining SDCD and PHOW MSDSIFT histograms into a one histogram. This fusion concatenates on the Homogeneous Kernel Map of SDCD histograms, mean of SDCD histograms, standard deviation of SDCD histograms, Homogeneous Kernel Map of PHOW histograms, mean of PHOW histograms, and standard deviation of PHOW histograms into one vector.
 5. SVM train by SVM Chi-square
 6. Extract scores from SVM
 7. Maximum pooling
 8. Testing of the model on previously-unseen pictures

To aid understanding, because this model has a lot of stages, we divided it in two parts, Figs. 2 and 3. Immediate results are captioned by a number of above stages. Figure 2 represents how color and PHOW MSDSIFT features are extracted. To test our model, both features (SDCD and PHOW MSDSIFT) are extracted and their histograms are generated. After that, both the color histogram and PHOW histograms change from nonlinear into linear by Homogeneous Kernel Map. Later, mean and standard deviation of both of the linear histograms are calculated and combined into a vector with 6 items: SDCD linear histogram, mean of linear SDCD histogram, standard deviation of linear SDCD histogram, PHOW linear histogram, mean of linear PHOW histogram, and standard deviation of linear PHOW histogram. Then, spatial histograms of visual words from test pictures were compared with spatial histograms of visual words based on SVM CHI^2 . With the help of scoring and maximum scores are pooled out. Afterward, the appropriate concept names were extracted by finding maximum score of retrieved concept names for test images. Figure 3 shows how

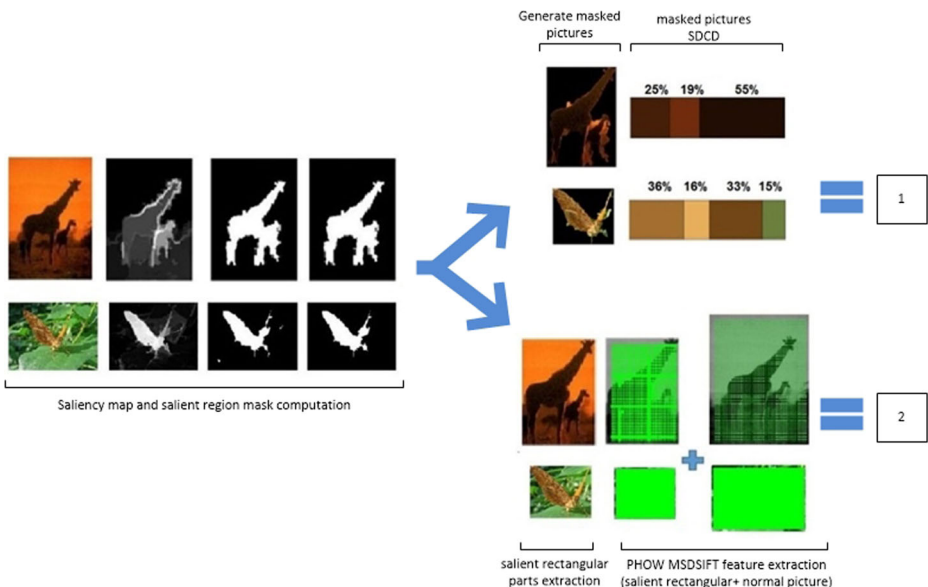


Fig. 2 SDCD and PHOW MSDSIFT BoVW model. Continued in the next Fig. 3

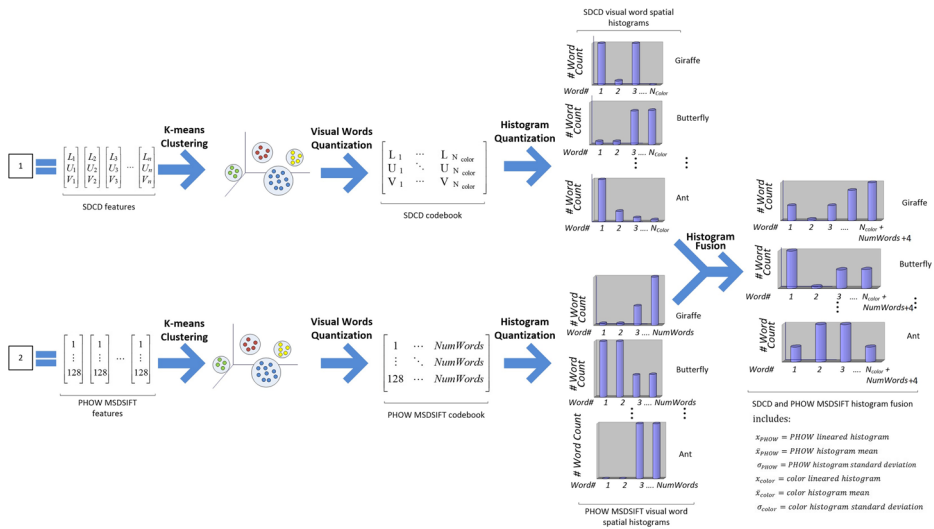


Fig. 3 SDCD& PHOW MSDSIFT BoVW model (second part). Followed after previous Fig. 2

visual words and histograms are quantized and, in the final stage, fused together by the proposed late fusion model. Fusion is the concatenation of both PHOW MSDSIFT and SDCD histograms, mean and standard deviation of each feature histograms’ Homogeneous Kernel Map. For more explanation internal fusion is described in the following.

3.3 The proposed model feature extraction and fusion in detail

In this subsection all the detail information about feature dimensions and explanations for immediate results are described in the following:

1. Images standardized for size to less than 300×300 pixels
2. Saliency map and salient region mask computation
3. SDCD BoVW stages:
 - (a) Generate masked pictures
 - (b) Extract SDCD of the masked pictures in LUV color space $3 \times C^N$, which C^N is the number of colors for all train pictures.
 - (c) Create SDCD codebook, and color table based on K-Means classification $N^{color} \times N^{train}$, and $N^{color} \times 3$ in which N^{color} is the number of colors remaining after SDCD codebook extraction and N^{train} is the number of train images.
 - (d) Quantize SDCD visual words spatial histograms, based on KD-trees $N^{train} \times N^{color}$
 - (e) Change SDCD histogram from non-linear into a compact linear by homogeneous kernel map.
4. PHOW MSDSIFT BoVW stages:
 - (a) Generate saliency rectangular parts of the picture
 - (b) Extract PHOW MSDSIFT feature from salient rectangular parts and normal pictures $128 \times N$, $128 \times M$ in which 128 is the dimension for SIFT features, N is the

- number of features of the salient rectangular part and M is the number of features of the normal picture. These features are combined into a bigger feature matrix with dimensions of $128 \times (M + 2 \times N)$. It means that the salient features are two times repeated and normal features just once.
- (c) Create PHOW MSDSIFT codebook, based on K-Means classification $128 \times N^{\#PHOW-codebook}$, which $N^{\#PHOW-codebook}$ is 1024 for Caltech-101 dataset and 2048 for Caltech-255 dataset.
 - (d) Quantize visual words spatial histograms, based on KD-trees $N^{train} \times N^{\#PHOW-codebook}$ in which N^{train} is the number of train images and
 - (e) Change PHOW histogram from non-linear into a compact linear by homogeneous kernel map.
5. Histogram fusion by combining SDCD, PHOW MSDSIFT histograms (x_{SDCD} , x_{PHOW}), mean (\bar{x}_{SDCD} , \bar{x}_{PHOW}), and standard deviation (σ_{SDCD} , σ_{PHOW}) of both of them into a one vector $N^{train} \times (N^{color} + N^{\#PHOW-codebook} + 4)$
 6. SVM train by SVM Chi-square.
 7. Extract SVM scores
 8. Maximum score pooling to recognize the object
 9. Testing of the model on previously-unseen pictures

4 Experimental setup

As mentioned earlier, this paper aims to investigate the potential and accuracy of the SDCD & PHOW MSDSIFT BoVW model for fusing SDCD color features and PHOW MSDSIFT features to recognize color objects in image retrieval. MSDSIFT scales are 4,6,8, and 10. MSDSIFT step (in pixels) of the grid at which the dense SIFT features are extracted is 2. Codebook is created based on elkan K-Means classification. Visual words quantize spatial histograms based on KD-trees. The proposed model is trained with SVM Chi-Square and scored histogram fusion is a concatenation of the linear SDCD histogram (x_{SDCD}), mean of SDCD histogram (\bar{x}_{SDCD}), standard deviation of SDCD histogram (σ_{SDCD}), the PHOW MSDSIFT SDCD histogram (x_{PHOW}), mean of PHOW MSDSIFT histogram (\bar{x}_{PHOW}), standard deviation of PHOW MSDSIFT histogram (σ_{PHOW}). The best result is selected by maximum pooling method. Evaluations are performed on the Caltech-101 dataset [32], in addition to the animal subset of the Caltech-256 [18] dataset. The number of codebooks are 1024 and 1500 for Caltech-101 and Caltech-256 respectively.

4.1 Caltech-101 dataset

This dataset has approximately 40–800 images per category. It contains a total of 9,146 images split between 101 distinct objects (including faces, watches, ants, pianos, etc.) and a background category (for a total of 102 categories). As suggested by Wang et al. [56] and other researchers [6, 18], the dataset is partitioned into 5, 10, ..., 30 training images per class and no more than 50 test images per class. The number of extracted code words was 1024. To make a comparison between this method and Vedaldi and Fulkerson [53] grayscale PHOW descriptor, the same training and test images were used. For comparison with color feature extraction methods, PHOW-color, the HSV histogram color, and RGB histogram color, presented by Vedaldi and Fulkerson [53], were used.

4.2 Caltech-256 dataset

From the Caltech-256 dataset, 20 different animals (bear, butterfly, camel, dog, house-fly, frog, giraffe, goose, gorilla, horse, humming bird, ibis, iguana, octopus, ostrich, owl, penguin, starfish, swan, and zebra) were selected from different environments (lake, desert, sea, sand, jungle, bushy, etc.). A common training setup (15, 30, 45, and 60 training images for each class) was followed [56]. There were less than 50 test images per class. To compare this method with the basic BoVW model, the same training and test images were chosen. The number of extracted code words was 1500.

4.3 Essential needs

The Essential software for running the program is: Matlab 2013a/2014a. The essential open source libraries for running the program are: Vlfeat: open source library implements popular computer vision algorithms specializing in image understanding and local features extraction and matching, and LIBSVM: A library for support vector machines.

4.4 Accuracy of proposed method

In the following section, we present the results obtained on the datasets and compare our method with two recent studies. For measuring accuracy, we used three famous methods: precision, accuracy, and classification rate which were also used in [6, 12, 17, 31, 52, 53]. Since these formulas are well known, we do not describe them in detail here.

5 Results and discussion

A comparison with Vedaldi's color descriptors (*PHOW + Opp.-MSDSIFT*, *PHOW + HSV-MSDSIFT*, and *PHOW+RGB-MSDDSIFT*) was done using the same train and test pictures and number of codebooks (Table 1). In this table, the proposed model consistently

Table 1 Comparison with other color descriptors results based on different number of codebook in Caltech-101 dataset

| Number of codebook & Method | 5 | 15 | 10 | 20 | 25 | 30 |
|-----------------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| 512 PHOW+Opp.MSDSIFT | 49.02 | 56.86 | 56.13 | 61.37 | 61.93 | 66.67 |
| 512 PHOW+HSV-MSDSIFT | 48.04 | 58.82 | 54.41 | 60.59 | 64.71 | 66.54 |
| 512 PHOW+RGB-MSDDSIFT | 55.88 | 64.05 | 62.75 | 68.43 | 70.92 | 72.06 |
| 512 SDCD & PHOW | 61.76 | 66.34 | 62.99 | 67.84 | 69.28 | 73.65 |
| 1024 PHOW+Opp.MSDSIFT | 49.02 | 57.19 | 58.82 | 62.94 | 65.03 | 67.65 |
| 1024 PHOW+HSV-MSDSIFT | 50 | 59.15 | 56.13 | 62.94 | 64.87 | 68.26 |
| 1024 PHOW+RGB-MSDDSIFT | 56.86 | 65.36 | 62.99 | 67.84 | 70.59 | 72.92 |
| 1024 SDCD & PHOW | 62.75 | 66.67 | 63.97 | 70.39 | 70.75 | 75.37 |
| 2048 PHOW+Opp.MSDSIFT | 48.04 | 57.19 | 59.07 | 64.51 | 64.05 | 67.52 |
| 2048 PHOW+HSV-MSDSIFT | 47.06 | 58.17 | 57.35 | 63.32 | 65.03 | 68.38 |
| 2048 PHOW+RGB-MSDDSIFT | 57.84 | 64.38 | 62.25 | 67.84 | 71.73 | 73.77 |
| 2048 SDCD & PHOW | 59.80 | 66.69 | 66.67 | 71.18 | 72.71 | 75.37 |

Table 2 Classification rate comparison based on percentage of classification rate in Caltech-101 dataset and different colored SIFT methods three states of arts (color SIFT [49], CSIFT [11], Color PHOW [53])

| Methods | Feature Combination | 5 | 10 | 15 | 20 | 25 | 30 | |
|--------------------|-----------------------------------|---|--------------|--------------|--------------|--------------|--------------|-------|
| Color SIFT | RGB-SIFT | | | 63.14 | | | 69.05 | |
| | Opp.SIFT | | | 50 | | | 58.72 | |
| | HSV-SIFT | | | 56.47 | | | 63.40 | |
| | Tranf.clrSIFT | | | 61.47 | | | 69.11 | |
| Color SIFT | RGB-SIFT+Opp.SIFT | | | 61.24 | | | 68.36 | |
| | RGB-SIFT+ Tranf.clrSIFT | | | 63.4 | | | 68.42 | |
| | RGB-SIFT+ HSV-SIFT | | | 61.63 | | | 69.28 | |
| | HSV-SIFT+ Tranf.clrSIFT | | | 62.16 | | | 70.30 | |
| | HSV-SIFT+ Opp.SIFT | | | 57.32 | | | 65.23 | |
| | Tranf.clrSIFT+ Opp.SIFT | | | 61.96 | | | 68.28 | |
| | Color SIFT | RGB-SIFT+ HSV-SIFT+ Opp.SIFT | | | 60.78 | | | 67.80 |
| Color SIFT | RGB-SIFT+ HSV-SIFT+ Tranf.clrSIFT | | | 63.53 | | | 69.39 | |
| | HSV-SIFT+ Opp.SIFT+ Tranf.clrSIFT | | | 61.63 | | | 69.25 | |
| | RGB-SIFT+ Opp.SIFT+ Tranf.clrSIFT | | | 62.81 | | | 69.20 | |
| | Color SIFT | HSV-SIFT+ Opp.SIFT+ Tranf.clrSIFT+ RGB-SIFT | | | 63.59 | | 69.92 | |
| | CSIFT | LLC + L^2 -norm+ YCbCr-SIFT | 47.18 | 57.39 | 62.41 | 65.98 | 68.17 | 69.74 |
| | Color PHOW | PHOW+ Opp.MSDSIFT | 49.02 | 57.19 | 58.82 | 62.94 | 65.03 | 67.65 |
| PHOW+ HSV-MSDSIFT | | 50 | 59.15 | 56.13 | 62.94 | 64.87 | 68.26 | |
| PHOW+ RGB-MSDDSIPT | | 56.86 | 65.36 | 62.99 | 67.84 | 70.59 | 72.92 | |
| Proposed | SDCD & PHOW | 62.75 | 66.67 | 63.97 | 70.39 | 71.75 | 75.37 | |

outperformed the other color descriptors under different number of train images using different numbers of codebooks; 1,024 codebooks always improve the final classification rate. Therefore, for the rest of the experiments, 1,024 codebooks is selected for the Caltech-101 dataset. The reason behind is the extraction of the dominant color extraction of the salient part instead of feature extraction from different color spaces. The proposed method for object classification performed 100% better than methods from three recent studies [11, 49, 53] and 19 different color feature extraction methods and different number of train images (5, 10, ..., 30) for the Caltech-101 dataset (see Table 2).

Table 3 The comparison of classification rate between SBBovW [41] and SDCD & PHOW MSDSIFT (the new model) in animal subset of Caltech-256

| No. of training images | 15 | 30 | 40 | 45 | 60 |
|------------------------|--------------|-----------|--------------|--------------|--------------|
| BoVW | 27.5 | 36.25 | 38.5 | 38.63 | 36.66 |
| SBBovW [41] | 35 | 46.88 | 51 | 55 | 50.33 |
| SDCD & PHOW (proposed) | 36.25 | 50 | 55.50 | 55.45 | 48.67 |






| | | | | | |
|-------------------|---|---|---|---|--|
| Picture Method |  |  |  |  |  |
| | BoVW | gorilla | owl | bear | starfish |
| SBBovW | gorilla | owl | gorilla | bear | gorilla |
| SDCD+PHOW | gorilla | gorilla | gorilla | gorilla | gorilla |

Fig. 4 Retrieved object name between BoVW, SBBovW, and SDCD + PHOW (proposed)

A comparison between the proposed model and the basic BoVW and previously proposed SBBovW model using the Caltech-256 dataset, is provided in Table 3 under different number of training images (15, 30, ..., 60). This table demonstrates that the proposed model performed 100% better than SBBovW model. This is due to the addition of SDCD color descriptor. Figure 4 is provided the retrieved object name for 5 test images between BoVW, SBBovW, and SDCD + PHOW (the proposed model). The proposed model retrieved 100% accurate names.

These results demonstrate the effectiveness of the proposed SDCD and SBBovW model for improving color object classification. The final accuracy and precision results are depicted in Figs. 5, 6 and 7 under the same train and test images. In Fig. 5, the final accuracy results are compared with the PHOW+MSDSIFT [53] and SBBovW for the Caltech-101 dataset. In addition, Fig. 6 shows the precision comparison between these (PHOW+MSDSIFT, SBBovW, and the proposed model) methods for the Caltech-101 dataset. This figure, showed that the proposed late fusion model outperforms the PHOW+MSDSIFT because of adding color and salient feature information for 56 concepts (*BACKGROUND-Google, Faces, Faces-easy, Leopards, Motorbikes, accordion, airplanes, anchor, binocular, butterfly, cellphone, chair, chandelier, cougar, crab, crocodile, cup, dollar-bill, dolphin, dragonfly, electric-guitar, euphonium, ferry, garfield, gerenuk, grand-piano, headphone, hedgehog, helicopter, ibis, inline-skate, joshua-tree, kangaroo, lamp,*

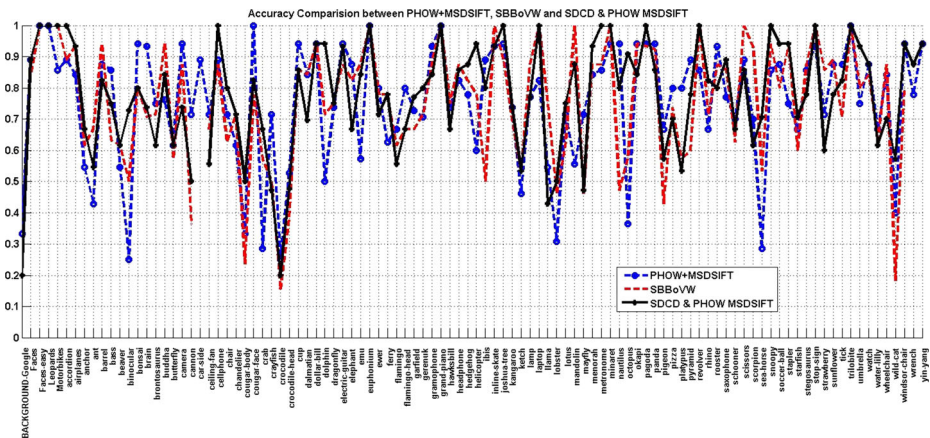


Fig. 5 Accuracy comparison of PHOW+MSDSIFT [53], SBBovW [41] and the proposed model on Caltech-101

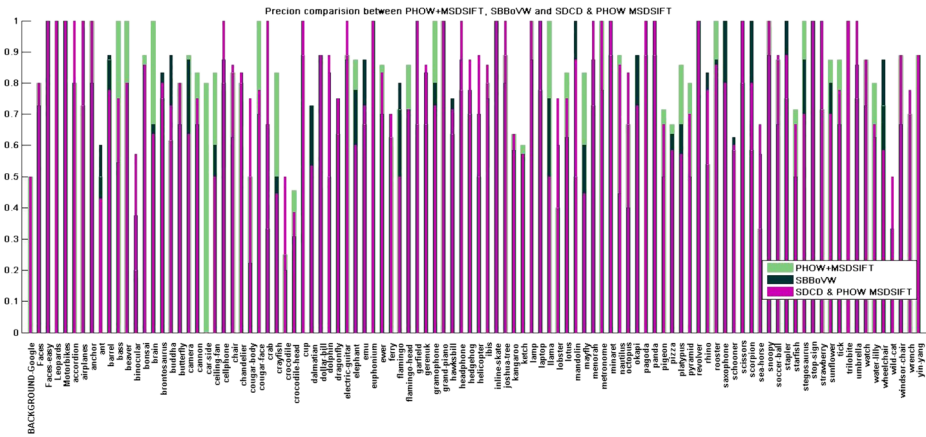


Fig. 6 Precision comparison of PHOW+MSDSIFT [53], SBBovW [41] and the proposed model on Caltech-101

laptop, lobster, menorah, metronome, minaret, octopus, pagoda, panda, revolver, scissors, sea-horse, snoopy, soccer-ball, stop-sign, strawberry, trilobite, umbrella, watch, wild-cat, windsor-chair, wrench, and yin-yang) but does not outperform for 46 concepts (*ant, barrel, bass, beaver, bonsai, brain, brontosaurus, buddha, camera, cannon, car-side, ceiling-fan, cougar-face, crayfish, crocodile-head, dalmatian, elephant, emu, ewer, flamingo, flamingo-head, gramophone, hawksbill, ketch, llama, lotus, mandolin, mayfly, nautilus, okapi, pigeon, pizza, platypus, pyramid, rhino, rooster, saxophone, schooner, scorpion, stapler, starfish, stegosaurus, sunflower, tick, water-lilly, and wheelchair*) due to incorrect extraction of

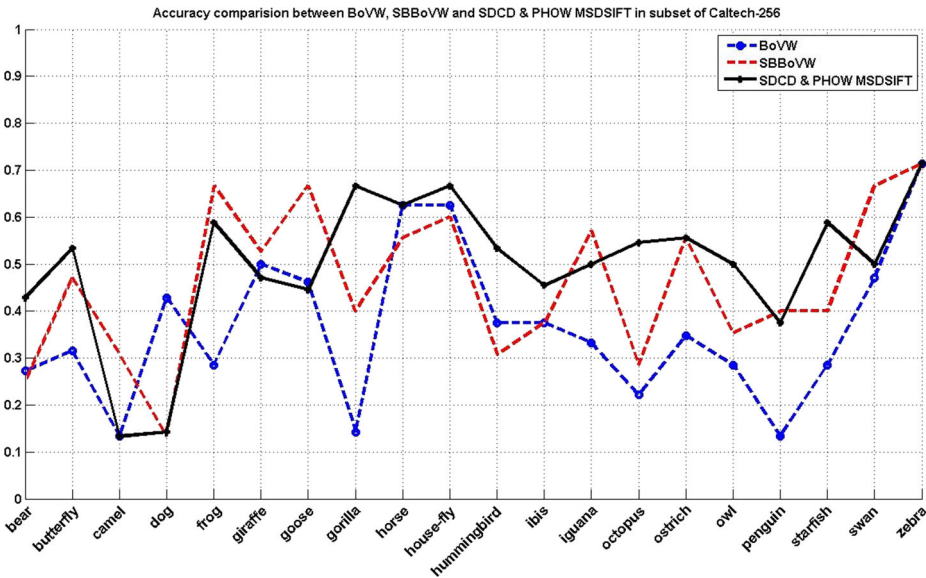


Fig. 7 Comparison of accuracy of BoVW model, SBBovW [41] model and proposed model on a subset of Caltech-256

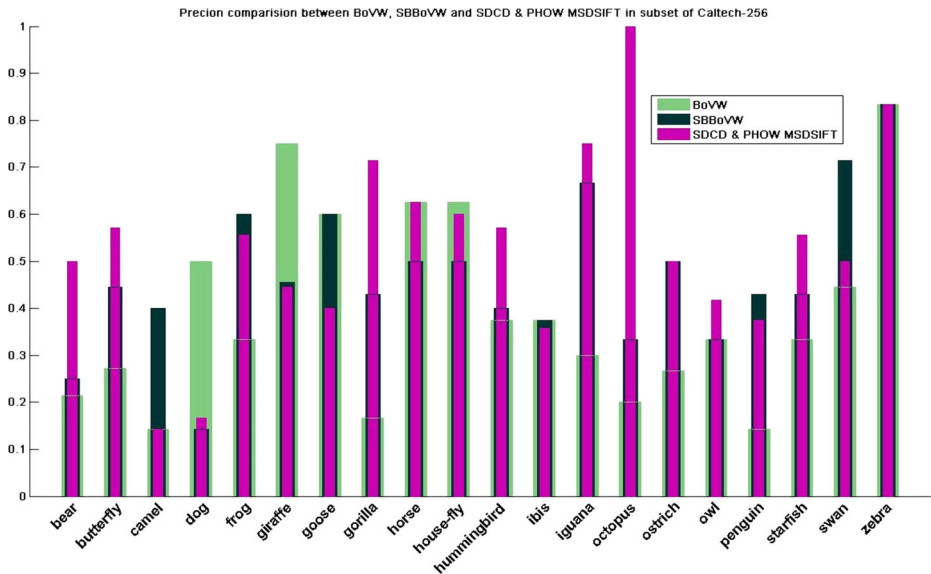


Fig. 8 Precision comparison of BoVW model, SBBovW [41] model and proposed model on a subset of Caltech-256

salient part for those objects which have narrow area lines, spotted or not smooth patterns or colors. These results were supported by the accuracy comparison (see Fig. 5).

In a detailed precision comparison of the proposed fusion model and BoVW model using the animal subset of the Caltech-256 dataset (see Fig. 8), the proposed late fusion model outperforms because of adding color and salient feature information for 11 concepts (*bear, butterfly, gorilla, horse, hummingbird, iguana, octopus, ostrich, owl, starfish, and zebra*) but does not outperform for those animals which do not have smooth patterns and colors (e.g. *giraffe, ...*) the salient extraction dose not work properly and gets worse results for 9 concepts (*camel, dog, frog, giraffe, goose, house-fly, ibis, penguin, and swan*). On the other hand, in comparison with the SBBovW method, the proposed late fusion model outperforms due to adding salient features and color information for 13 concepts (*bear, butterfly, dog, gorilla, horse, house-fly, hummingbird, iguana, octopus, ostrich, owl, starfish, and zebra*) but does not outperform for seven concepts, because of incorrect salient object extraction on animals which have spotted patterns or not smooth colors (e.g. *camel, frog, giraffe, goose, ibis, penguin, and swan*). These results are supported by the accuracy comparison for the PHOW+MSDSIFT [53], the previously proposed SBBovW, and the proposed model in Caltech-101 dataset (see Fig. 7).

Based on these results, the proposed fusion model improves the final precision, accuracy, and classification rate in images in which the salient region could be correctly extracted.

6 Conclusions and future research

In this paper, first a new SDCD algorithm to extract colors of the salient object of a picture was presented. Using this algorithm, a new model, SDCD & PHOW MSDSIFT BoVW, was proposed to fuse SDCD histogram with PHOW MSDSIFT histogram. The proposed model

classifies color objects that have had low accuracy in prior methods. This method mixes SDCD and PHOW MSDSIFT features of the original and salient parts of pictures and fuses them with a new fusion model together to generate better codebooks. The final results and comparison with 3 state-of-the-art models and 19 different color feature extraction methods shows that extraction of SDCD colors improved the final results. However, this model still needs improvements for those objects for which color is not as effective in their classification. In the future, with the help of other features, such as texture, difficult objects can be recognized more accurately. In addition, multi-object datasets, such as VOC-7, may present another approach to improve the proposed model. Parallel processing is another future area to run the code faster.

References

1. Abdelkhalak B, Zouaki H (2015) Content-based bird retrieval using shape context, color moments and bag of features. *Int J Comput Sci Issues (IJCSI)* 12(1):101
2. Alqasrawi Y, Neagu D, Cowling PI (2011) Fusing integrated visual vocabularies-based bag of visual words and weighted colour moments on spatial pyramid layout for natural scene image classification. *SIViP* 7(4):759–775
3. Bannour H, Hudelot C (2013) Building and using fuzzy multimedia ontologies for semantic image annotation. *Multimedia Tools Appl* 72(3):2107–2141
4. Barata C, Marques JS, Rozeira J (2006) Evaluation of color based keypoints and features for the classification of melanomas using the bag-of-features model. In: *Proceedings of the 9th European conference on computer vision computer vision – ECCV 2006, Part I. Graz*, pp 40–49
5. Bay H, Tuytelaars T, Van Gool L (2006) SURF: Speeded up robust features, pp 404–417
6. Berg AC (2006) SVM-KNN: Discriminative nearest neighbor classification for visual category
7. Borji A (2014) What is a salient object? A dataset and a baseline model for salient object detection. (Xxx):1–15
8. Borji A, Sihite DN, Itti L (2012) Salient object detection: a benchmark, pp 414–429
9. Borji A, Cheng M-M, Jiang H, Li J (2014) Salient object detection: a survey, pp 1–26
10. Bosch A, Zisserman A, Mu X, Munoz X (2007) Image classification using random forests and ferns. *Iccv*, pp 1–8
11. Chen J, Li Q, Peng Q, Wong KH (2015) Csift based locality-constrained linear coding for image classification. *Pattern Anal Appl* 18(2):441–450
12. Chiang C-C (2013) Interactive tool for image annotation using a semi-supervised and hierarchical approach. *Comput Standards Interfaces* 35(1):50–58
13. Csurka G, Dance CR, Fan L, Willamowski J, Bray C, Maupertuis D (2004) Visual categorization with bags of keypoints. In: *Workshop on statistical learning in computer vision, ECCV, vol 1*, pp 1–2
14. Dalal N, Triggs B, Europe D (2005) Histograms of oriented gradients for human detection
15. Deng Y, Kenney C, Moore MS, Manjunath BS (1999) Peer group filtering and perceptual color image quantization IV-22. In: *Proceedings of the 1999 IEEE international symposium on circuits and systems, 1999. ISCAS'99, vol 4*, pp 21–24
16. Dey V, Zhang Y, Zhong M, Geomatics Engineering (2010) A review on image segmentation techniques with. XXXVIII:31–42
17. Fakhari A, Moghadam AME (2013) Combination of classification and regression in decision tree for multi-labeling image annotation and retrieval. *Appl Soft Comput* 13(2):1292–1302
18. Griffin P, Holub G, Perona A (2007) Caltech-256 object category dataset
19. Hua G, Wang L, Xue J, Zheng N (2011) Automatic salient object extraction with contextual cue. In: *2011 international conference on computer vision*, pp 105–112
20. Ionescu RT, Popescu M, Grozea C (2007) Local learning to improve bag of visual words model for facial expression recognition
21. Islam M, Zhang D, Lu G (2008) Automatic categorization of image regions using dominant color based vector quantization. In: *Proceedings - digital image computing: techniques and applications, DICTA 2008*, pp 191–198
22. Jalali S, Tan C, Ong S-H, Seekings PJ, Taylor EA (2013) Visual recognition using a combination of shape and color features. In: (CogSci), the annual meeting of the cognitive science society, pp 2638–2643

23. Jiang H, Wang J, Yuan Z, Liu T, Zheng N, Li S (2011) Automatic salient object segmentation based on context and shape prior. In: *BMVC*, vol 6, p 9
24. Jiang H, Wang J, Yuan Z, Wu Y, Zheng N, Li S (2013) Salient object detection: a discriminative regional feature integration approach. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, pp 2083–2090
25. Khan FS, vande Weijer J, Vanrell M (2011) Modulating shape features by color attention for object recognition. *Int J Comput Vis* 98(1):49–64
26. Kim J, Grauman K (2011) Boundary preserving dense local regions. In: *Cvpr 2011*, pp 1553–1560
27. Kim M-U, Yoon K (2014) Performance evaluation of large-scale object recognition system using bag-of-visual words model. *Multimedia Tools and Applications*
28. Kulkarni G, Premraj V, Dhar S, Li S, Choi Y, Berg AC, Berg TL (2011) Baby talk: understanding and generating simple image descriptions. *Cvpr 2011*, pp 1601–1608
29. Lampert CH, Blaschko MB, Hofmann T (2008) Beyond sliding windows: object localization by efficient subwindow search. In: *2008 IEEE conference on computer vision and pattern recognition*, pp 1–8
30. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: *2006 IEEE computer society conference on computer vision and pattern recognition - volume 2 (CVPR'06)*, vol 2, pp 2169–2178
31. Lee C-H, Yang H-C, Wang S-H (2011) An image annotation approach using location references to enhance geographic knowledge discovery. *Expert Syst Appl* 38(11):13792–13802
32. Li F-F, Fergus R, Perona P (2007) Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Comput Vis Image Underst* 106(1):59–70
33. Liu C, Yuen J, Torralba A, Sivic J, Freeman WT (2008) SIFT flow: dense correspondence across different scenes. *1(1)*:28–42
34. Liu W, Tao D (2013) Multiview hessian regularization for image annotation. *IEEE Trans Image Process* 22(7):2676–2687
35. Liu W, Tao D, Cheng J, Tang Y (2014) Multiview Hessian discriminative sparse coding for image annotation. *Comput Vis Image Underst* 118:50–60
36. Long X, Lu H, Li W (2012) Image classification based on nearest neighbor basis vectors. *Multimedia Tools Appl* 71(3):1559–1576
37. Lowe DG (1999) Object recognition from local scale-invariant features. In: *Proceedings of the seventh IEEE international conference on computer vision*, vol 2, pp 1150–1157
38. Luo Y, Liu T, Tao D, Xu C (2015) Multiview matrix completion for multilabel image classification. *IEEE Trans Image Process* 24(8):2355–2368
39. Luo Y, Wen Y, Tao D, Gui J, Xu C (2016) Large margin multi-modal multi-task feature extraction for image classification. *IEEE Trans Image Process* 25(1):414–427
40. Mansourian L, Abdullah MT, Abdullah LN, Azman A (2015) Evaluating classification strategies in bag of sift feature method for animal recognition. *Res J Appl Sci Eng Technol* 10(11):1266–1272
41. Mansourian L, Abdullah MT, Abdullah LN, Azman A, Mustaffa MR (2016) A salient based bag of visual word model (sbbvov): Improvements toward difficult object recognition and object location in image retrieval. *KSII Trans Internet Inf Syst* 10(2):769–786
42. Mesleh AMA (2007) Chi square feature extraction based svms arabic language text categorization system. *J Comput Sci* 3(6):430–435
43. Mikolajczyk K, Schmid C (2004) Scale & affine invariant interest point detectors. *Int J Comput Vis* 60(1):63–86
44. Mikolajczyk K, Leibe B, Schiele B, Darmstadt TU (2005) Local features for object class recognition
45. Murphy K, Torralba A, Eaton D, Freeman W (2006) Object detection and localization using local and global features, pp 382–400
46. O'Hara S, Draper BA (2011) Introduction to the bag of features paradigm for image classification and retrieval, pp 1–25
47. Oquab M (2012) Is object localization for free? Weakly-supervised learning with convolutional neural networks. (iii)
48. Li F-F, Perona P (2005) A Bayesian hierarchical model for learning natural scene categories. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol 2, pp 524–531
49. Rassem TH, Khoo BE (2011) Object class recognition using combination of color sift descriptors. In: *2011 IEEE international conference on imaging systems and techniques (IST)*. IEEE, pp 290–295
50. Talib A, Mahmuddin M, Husni H, George LE (2013) A weighted dominant color descriptor for content-based image retrieval. *J Vis Commun Image Represent* 24(3):345–360

51. Tian D (2014) Semi-supervised learning for automatic image annotation based on bayesian framework. *Intern J Control Autom* 7(6):213–222
52. Tousch AM, Herbin S, Audibert JY (2012) Semantic hierarchies for image annotation: a survey. *Pattern Recog* 45(1):333–345
53. Vedaldi A, Fulkerson B (2010) VLFeat - an open and portable library of computer vision algorithms. *Design* 3(1):1–4
54. Vigo DAR, Khan FS, van de Weijer J, Gevers T (2010) The impact of color on bag-of-words based object recognition. In: 2010 20th international conference on pattern recognition, pp 1549–1553
55. Wang H, Nie F, Huang H (2013) Multi-view clustering and feature learning via structured sparsity. In: Proceedings of the 30th international conference on machine learning (ICML-13), pp 352–360
56. Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y Locality-constrained linear coding for image classification
57. Wang P, Wang J, Zeng G, Feng J, Zha Hongbin, Li S (2012) Salient object detection for searched web images via global saliency. In: 2012 IEEE conference on computer vision and pattern recognition, pp 3194–3201
58. Van De Weijer J, Khan FS (2013) Fusing color and shape for bag-of-words, pp 25–34
59. Yamada A (2001) MPEG-7 Visual part of experimentation Model Version 9.0. ISO/IEC JTC1/SC29/WG11/N3914
60. Yan Q, Xu L, Shi J, Jia J (2013) Hierarchical saliency detection. In: 2013 IEEE conference on computer vision and pattern recognition, pp 1155–1162
61. Yang N, Kuo C, Chang W, Lee T (2008) A fast method for dominant color descriptor with new similarity measure. *iscom2005*
62. Zhang D (2004) Improving image retrieval performance by using both color and texture features. In: 3rd international conference on image and graphics (ICIG'04), pp 4–7
63. Zhang D, Islam MM, Lu G (2012) A review on automatic image annotation techniques. *Pattern Recog* 45(1):346–362
64. Zhang D, Islam MM, Lu G (2013) Structural image retrieval using automatic image annotation and region based inverted file. *J Vis Commun Image Represent* 24(7):1087–1098
65. Zhang J, Marszaek M, Lazebnik S, Schmid C (2006) Local features and kernels for classification of texture and object categories: a comprehensive study. *Int J Comput Vis* 73(2):213–238
66. Zhong S-h, Liu Y, Liu Y, Fu-lai C (2012) Region level annotation by fuzzy based contextual cueing label propagation. *Multimedia Tools Appl* 70(2):625–645



Leila Mansourian is currently a Senior Lecturer at the Faculty of Computer Science and Information Technology, Khayyam Higher Education University, Mashhad (KHEU), and Islamic Azad University, Quchan Branch (IAUQ) in Iran. She received her PHD in Information Retrieval from Faculty of Computer Science and Information Technology at Universiti Putra Malaysia, in 2016, and M.S. degree in Artificial intelligence from Islamic Azad University, Science and Research Branch, Tehran, in 2007, and Bac. in Computer Science from Islamic Azad University of Mashhad, in 2002. Her research interests include Multimedia Systems and Applications, Multimedia Information Retrieval and Pattern Recognition.



Muhamad Taufik Abdullah is an Associate Professor at the University Putra Malaysia. He received his BS degree in Computer Science from the Universiti Pertanian Malaysia in 1990, his MS degree in Computer Science from the Universiti Teknologi Malaysia in 1992, and his PhD degree in Information Retrieval from the Universiti Putra Malaysia in 2006. He is the author of more than 20 journal papers and has written two book chapters. His current research interests include information retrieval, natural language processing, and multimedia computing.



Lili Nurliyana Abdullah is currently an Assistant Professor in the Department of Multimedia, Faculty of Computer Science and Information Technology at Universiti Putra Malaysia. She received her Ph.D degree in Information Science from Universiti Kebangsaan Malaysia, in 2007, M.S. degree in Engineering (Telematics) from University of Sheffield, United Kingdom, in 1996, and Bac. in Computer Science from UPM in 1990. Her research interests include multimedia system, video processing and retrieval, computer modelling and animation, image processing and computer games.



Azreen Azman is a Senior Lecturer at the Universiti Putra Malaysia. He received a Diploma in Software Engineering from the Institute of Telecommunication and Information Technology in 1997. Immediately, he was accepted directly to second year in Multimedia University, Malaysia to study Bachelor of Information Technology majoring in Information Systems Engineering. He completed his bachelor degree in 1999. After serving in the industry for a few years, he enrolled for a Ph.D in January 2003, studying Computing Science specializing in Information Retrieval in the University of Glasgow, Scotland and completed his study in September 2007. His current research interests include information retrieval, text mining, opinion mining and semantic technology.



Mas Rina Mustaffa is currently a Senior Lecturer at the Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She received her BCompSc degree (Multimedia) and MSc degree (Multimedia Systems) in 2003 and 2006 respectively. She received her PhD for studies in Content-based Image Retrieval (CBIR) in 2012. All of the three degrees are from Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia. She has authored several publications in various journals and proceedings and presented at many conferences. She also has been actively involved in several international conferences as technical program committee. She is a member of IEEE, ACM, Malaysian Society of Information Retrieval and Knowledge Management (PECAMP), and International Association of Computer Science and Information Technology (IACSIT). Her primary research interests are multimedia systems and applications, Content-Based Image Retrieval (CBIR), image processing, pattern recognition, and interactive multimedia.