2020

Adaptive bag-of-visual word modelling using stacked-autoencoder and particle swarm optimisation for the unsupervised categorisation of images

Abass Olaode
University of Wollongong, aao808@uowmail.edu.au

Golshah Naghdy
University of Wollongong, golshah@uow.edu.au

Adaptive bag-of-visual word modelling using stacked-autoencoder and particle swarm optimisation for the unsupervised categorisation of images

Abstract

© The Institution of Engineering and Technology 2020 The bag-of-visual words (BOVWs) have been recognised as an effective mean of representing images for image classification. However, its reliance on a visual codebook developed using handcrafted image feature extraction algorithms and vector quantisation via k-means clustering often results in significant computational overhead, and poor classification accuracies. Therefore, this study presents an adaptive BOVW modelling, in which image feature extraction is achieved using deep feature learning and the amount of computation required for the development of visual codebook is minimised using a batch implementation of particle swarm optimisation. The proposed method is tested using Caltech-101 image dataset, and the results confirm the suitability of the proposed method in improving the categorisation performance while reducing the computational load.

Disciplines

Engineering | Science and Technology Studies

Publication Details

A. Olaode & G. Naghdy, "Adaptive bag-of-visual word modelling using stacked-autoencoder and particle swarm optimisation for the unsupervised categorisation of images," IET Image Processing, vol. 14, (9) pp. 1769-1776, 2020.

Adaptive Bag-of-Visual Word Modelling using Stacked-Autoencoder and Particle Swarm Optimisation for the Unsupervised Categorisation of Images

Abass Olaode [1*], Golshah Naghdy [2]

[1,2] School of Electrical Computer and Telecommunication Engineering, University of Wollongong, NSW, Australia
[*] Abass.Olaode808@uowmail.edu.au

Abstract: The Bag-of-Visual Words has been recognised as an effective mean of representing images for image classification. However, its reliance on a visual codebook developed using Hand Crafted image feature extraction algorithms and vector quantisation via k-means clustering often results in significant computational overhead, and poor classification accuracies. Therefore, this paper presents an adaptive Bag-of-Visual Word Modelling in which Image Feature Extraction is achieved using Deep Feature Learning and the amount of computation required for the development of Visual Codebook is minised using a batch implementation of Particle Swarm Optimisation. The proposed method is tested using Caltech 101 image dataset, and the results confirm the suitability of the proposed method in improving the categorisation performance while reducing the computational load.

1. Introduction

The semantic based annotation of images has been recognised as a viable means of bridging the semantic gap associated with Content Based Image Retrieval (CBIR) [1, 2, 3, 4, 5, 6]. While the efficient annotation of a large image collection via supervised machine learning remains a challenge in computer vision and image retrieval [7, 8, 9], the application of Unsupervised Machine Learning principles such as K-means clustering, Self-Organising Maps or Hierarchical clustering [10, 11] enables the image models computed from a given a large image collection to be grouped based on similarity [12, 13, 14, 15], without the need for labelled training samples, therefore is a natural fit for achieving Image annotation [16, 17, 18]. However, to achieve such unsupervised categorisation, there is a need for an efficient and effective local image pattern representation and global image representations [19]. This paper presents an unsupervised image categorisation built on Bag-of-Visual words (BOVW) image modelling of images as a suitable means of achieving efficient global representation of images for effective large-scale annotation.

The BOVW model of an image represents the image with a histogram showing the number of times the visual words belonging to a BOVW codebook appears on the image [20, 21, 22, 23] and has been popular in recent image classifications work [9]. However, the codebook development stage of BOVW modelling has been identified as a very computationally expensive stage because of the need to handle a very large number of features extracted from images belonging to the collection to be classified [21, 24, 25].

Furthermore, the number of visual words in a BOVW codebook has a direct influence on the dimensionality of image BOVW models, and determines how fast and accurate the image classification process will be [22, 26, 27]. If the number of Visual Words present in the resulting BOVW Codebook is not optimized for the image collection to be classified, the dimensionality of the image BOVW representation can become unnecessary long, thus making the classification process inefficient and the accuracies yielded will be lower than possible or the dimensionality can be too short for a reliable classification process [22, 26, 27]. Therefore, this paper identify the image feature extraction and vector quantisation stages as the two sub-stages of the BOVW modelling process that can be modified for performance optimisation of the image categorisation process, demonstrates the benefits of using deep feature learning as the image feature extraction algorithm in BOVW modelling and presents vector quantisation via a batch implementation of Particle Swarm Optimisation (PSO) as a means of achieving an efficient BOVW modelling of images.

The remainder of this paper is structured as follows: Section II provides a detailed discussion on recent research developments in Image representations using BOVW and Deep Feature Learning, while Section III describes the proposed adaptive image modelling approach. Section IV describes the experimental implementation and evaluation of the proposed algorithm through its application in the unsupervised classification of an image dataset. Section V analyses the experimental results, by showcasing the improvements in accuracies demonstrated by the proposed approach compared to existing methods. Section VI concludes the paper with a summary of the performance of the proposed algorithm in codebook development and its applicability in the semantic labelling of images.

2. Related works

Although Global Image representation via the Bag of Visual Word (BOVW) has been popular over the last two decades [28, 29, 30, 31, 32, 33, 34, 35], and has been recognised to be most appropriate for Unsupervised Image categorisation process [20, 36, 37], the need to quantise a large number of image features into Visual Words using the K-Means algorithm during the BOVW codebook development creates a heavy a number of computational problems [21, 24, 25, 38, 39], and often yields Visual Words that do not guarantee optimum classification performance. Therefore, towards reducing the number of image features to be handled during BOVW Codebook Development and to allow, this section reviews some

previous works related to the application of Deep Feature Learning to Image Representation and Vector quantisation in BOVW Image modelling.

2.1. *The application of deep feature learning to image pattern representation*

Deep Feature Learning has been recognised in image retrieval researches as a reliable method for generating a high-level image representation from a massive collection of images [16, 16, 40, 41, 42, 43, 44], and has been found to be an important inclusion in the implementation of automatic image annotation due to its strong discriminatory power of Deep Learning Image representations [41, 45].

A typical implementation of deep learning algorithm employs multiple layers of Machine Learning such as Independent Component Analysis (ICA), Convolutional Neural Network (CNN) and Stacked-Autoencoders where each layer receives its input from a previous layer [43], and the image representation is generated at the final layer.

While the global image representation via deep feature learning has become popular in computer vision and image retrieval researches [46, 47, 48, 49, 50, 51, 52, 53], such application of deep feature learning requires supervised fine-tuning as shown in Figure 1 for optimum image classification performance [54, 55, 56], therefore not readily suitable for unsupervised Image Categorisation. However, the Autoencoder; a popular algorithm for the implementation of deep feature learning, has been recognised to be more efficient than other manifold learning for the purpose of non-linear dimension reduction [47], a characteristic that makes it suitable for the development of local image pattern representation, where supervised fine-tuning is not necessary therefore can support a completely unsupervised image classification.

In [8] the authors demonstrated that the opportunity to change the number of layers and the number of neurons in each layer of a Deep Learning algorithm allows the feature extraction process to be adaptable to the content diversity of the image collection during BOVW modelling, thus generating image feature vectors whose dimension guarantees optimum discrimination, unlike the fixed 128 dimensions of Scale Invariant Feature Transform (SIFT) and 64 dimensions of Speeded-Up Robust Feature (SURF) [57, 58, 59].

The results shown in [8] confirms the applicability of image feature extraction via Stacked-Autoencoder to the BOVW modelling process. Although unlike SIFT, Stacked-Autoencoder (and other Deep Feature Learning algorithms) do not provide scale and rotation Invariance representations [60], the results in [8] confirms that this deficiency is largely compensated for by the histogram representation approach of BOVW and the spatial pyramid included in the image modelling for the elimination of spatial incoherency.

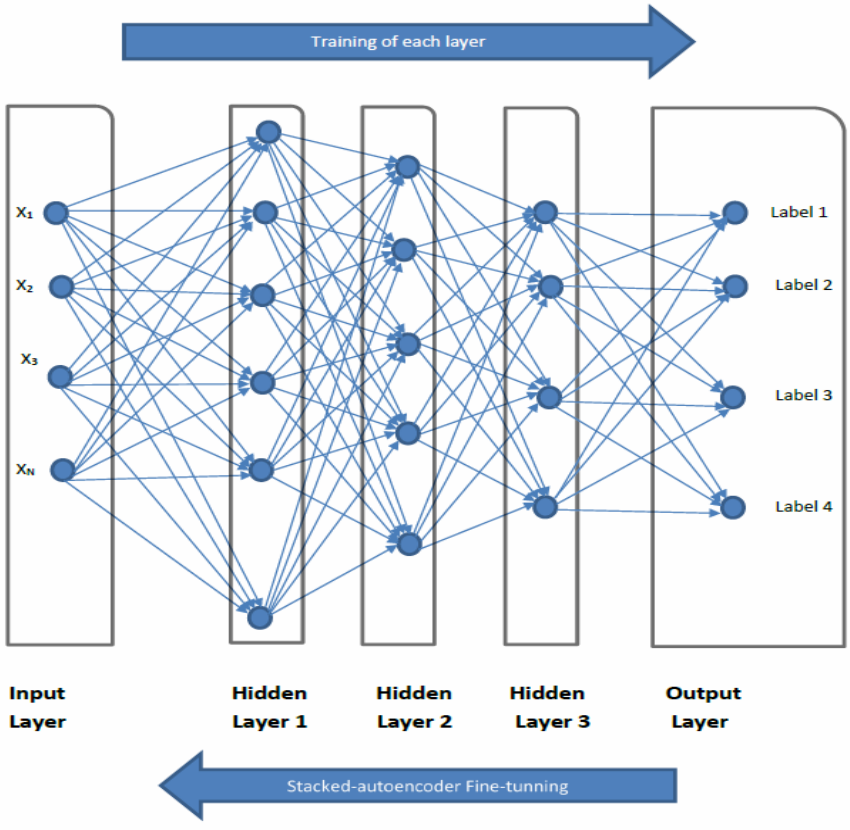

Fig 1. *A Stacked Autoencoder Network for Image Classification showing the Unsupervised Training Phase and the Supervised Fine-tuning phase*

2.2. Vector Quantisation in BOVW Codebook Development

In its simplest form, the codebook development stage of the BOVW image modelling is typically achieved by clustering available image features into a chosen distinct number of groups, after which the centroids of these groups are taken as the quantisation levels. An important advantage of the application of Deep Feature learning at this stage is the opportunity to control the number of image features to be collected from each image in the collection to be processed thus avoiding excessive computational overhead, commonly associated with sparse image features such as SIFT or SURF where the number of image features per image is not pre-determined or in Dense-SIFT where the number of features per image can be more than 10,000 with no means of controlling the number of image features.

The most popular method for achieving the required vector quantisation during BOVW codebook development is the K-means algorithm [20, 21, 22]. In the K-mean algorithm the centroid is the result of several attempts (iterations) aimed at minimising an overall measure of cluster quality (the objective function) [61]. However, there is need for other alternatives because of the tendency of the K-means algorithm to converge to wrong centers, especially due to the large number of image features typically generated during BOVW codebook development which also causes the vector quantisation via the K-mean algorithm to be a computationally intensive [20, 62].

Jurie and Triggs [63] demonstrated that the use of K-means clustering in development of BOVW codebooks is mainly reliable for handling homogenous image collections but is not adequate for handling natural object recognition tasks because the latter's statistics are less uniform [63]. Tirilly et al. [21] explain that attempts at speeding up the process by replacing K-means clustering with approximate algorithms often results in noisy visual words [21].

Although in the attempt to boost the categorisation of a BOVW process, Wu *et al.* [22] retained the K-Means algorithm in their proposed BOVW vector quantisation algorithm, while replacing Euclidean distance; the traditional image feature similarity measure with the Histogram Intersection Kernel (HIK). However, the accuracies obtained with the application of this codebook approach is only 2% to 4% better than the accuracies obtained with traditional approach, while incurring a significant increment in the computational time needed to complete the codebook development process, thus making the approach inefficient and unsuitable for handling large number of images. Therefore, there is a need to identify a suitable replacement that will guarantee good accuracy with minimum computational overhead.

Another drawback of vector quantisation via K-mean clustering is that the number of quantisation levels needs to be known at the beginning of the quantisation process [64, 65, 66]. Arbitrarily choosing a small codebook size may limit the classification process's discriminative power [27], while a larger than necessary codebook size will incur surplus processing overhead [26, 67]. Although Tsai [62] recommended a codebook size of 1000 visual words, the authors explained that the number of visual-words is dependent on the dataset [62]. Guo *et al.* [25] also explain that classification performance usually improves as the Visual codebook size initially increases, but it begins to deteriorate as the codebook size becomes larger [25]; thus confirming the need to pick a BOV codebook size that is adequate for the image collection being classified.

In the effort to improve the performance of Bag of Visual Phrase, Battiato *et al.* [68] recognised that better results could be achieved through the inclusion of a step that exploits the nature of the feature spaces during the codebook generation. Such strategy implemented in the visual codebook approach proposed in [69], where the process determined the appropriate number of visual words needed in a codebook by using a pseudo clustering algorithm [70] to eliminate repeated visual words from an available visual word set.

The X-Mean algorithm proposed by Pelleg and More [71] is a clustering algorithm designed for overcoming the need for the number of clusters to be specified at the beginning of a clustering process. Starting with an assumed minimal number of clusters, the X-mean algorithm implements the K-Means clustering repeatedly with an increasing number of clusters K, while measuring each of the clustering performance using Bayesian Information Criterion (BIC) until an assumed maximum number of clusters is reached. At the end of the clustering process, the value of K with the best clustering performance is then chosen as the appropriate value.

Although the X-Means algorithm was successfully applied to the BOVW Codebook Development by Kersorn *et al.* [72], the X-means method of implementing clustering several time in the search for the appropriate number of cluster is a computationally expensive process, when the number of image features to be quantised is large and each of the features are represented with high dimensional vectors (50 dimensions and above). Furthermore, the X-Means implementation does not include an explicit method of avoiding the problem of clustering process converging to wrong centers. Therefore, there is a need to further explore the behavior of X-Mean Clustering.

Recently, the application of PSO for data clustering has become popular [73, 74, 75, 76]. The PSO algorithm applies animal group information sharing behaviour to solving learning problems in a large data space [73]. Given a set of data samples X, represented as positions in a multi-dimensional space, the PSO algorithm attempts to identify best positions to represent the distribution of the samples within the multi-dimensional space. Where x_i, v_i, and y_i are the current position, current velocity, and the best position found so far for a particle p_i, the particle's position can be changed in accordance with Equation 1 and Equation 2 [74].

$$v_{i,k}(t+1) = wv_{i,k}(t) + c_1 r_{1,k}(t)\left(y_{i,k}(t) - x_{i,k}(t)\right) + c_2 r_{2,k}(t)(y(t) - x_{i,k}(t)) \quad (1)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (2)$$

Where W is the inertia weight, c_1 and c_2 are acceleration constants, and r_1 and r_2 are samples from a uniform distribution [74]. Equation 1 and Equation 2 are repeated in iterations, while the best position is determined using Equation 3 as shown below [74];

$$y_i(t+1) = \begin{cases} y_i(t) & if \quad f\big(x_i(t+1)\big) \geq f(y_i(t)) \\ x_i(t+1) & if \quad f\big(x_i(t+1)\big) < f(y_i(t)) \end{cases} \tag{3}$$

Although the search for optimum clustering solutions using this population-based search approach of PSO has proven to yield better result than K-Means [74], the traditional implementation of PSO does not provide an explicit method on how to pick initial solutions, and the commonly used approach of randomly picking initialisation particles from the a set containing thousands of samples as it is done in K-Means clustering (other K-means based clustering) exposes the process to convergence to dead centers or division of a single cluster into multiple clusters.

Also, like the K-Means clustering algorithm the PSO clustering process does not include the determination or how to pick the appropriate number of clusters. Therefore, towards the implementation of PSO without a prior knowledge of the number of inherent groups, this study presents an initialisation process that present a surplus number of seeds from which only the seeds that attract adequate number of samples are selected, thereby solving the both the initialisation problem and the determination of the number of clusters.

Furthermore, for efficient handling of large number of image features during the BOVW codebook development process, this study also present batch vector quantisation. Section III describes the novel steps are integrated into the proposed adaptive BOVW Modelling.

3. The proposed method

In general, the BOVW Codebook development process can be divided into two stages; the extraction of image features and the quantisation of the extracted image features into Visual words. This section provides a detailed description of the implementation of Image feature extraction via a 3-Layered Stacked-Autoencoder, and the batch vector quantisation process which uses PSO to generate visual words needed for the development of image BOVW representation that adequately considers the semantic content of the images to be classified, and to ensure good classification accuracy while minimising computational overhead.

3.1. Image feature extraction using stacked-autoencoder

One of the main reasons for the high computational overhead of vector quantisation via the K-means algorithm is the massive amount features generated from each image [72] especially when using dense feature extraction algorithm. The number of image features obtained from an image can be significantly reduced by taking advantage of the spatial redundancy of images [69], and limiting image feature extraction to evenly spaced locations within the image space by dividing the image into tiles using a moving window centered on evenly spaced locations within the image space.

While both overlapping and non-overlapping spatial tiling has been demonstrated to be effective in this regard [69], dividing the image into overlapping tiles facilitates an exhaustive search for content objects during feature extraction thereby supporting object recognition while still limiting the features obtainable from the image to the chosen number. Figure 2 is an illustration of tiles obtained from a sample image of a Leopard chosen from the Caltech-101 Objects Categories.

All the image tiles obtained from an experimental image collection are used to train the Deep Feature Learning Algorithm, after which the rows of each tile are concatenated to yield a single vector which is applied to the input layer of the trained Deep Learning algorithm to produce a corresponding image feature representation [46].

In [8], the authors demonstrated that Stacked-Autoencoder image feature extraction's approach of reducing the number of features by taking advantage of the spatial redundancy during the spatial tiling resulted in considerable reduction in the categorisation time when compared to SIFT. Although the time taken is higher than the time taken to complete the unsupervised categorisation with SURF features due to the time taken to train the Stacked-Autoencoder, the higher accuracy recorded by Stacked Autoencoder confirms its better efficiency.

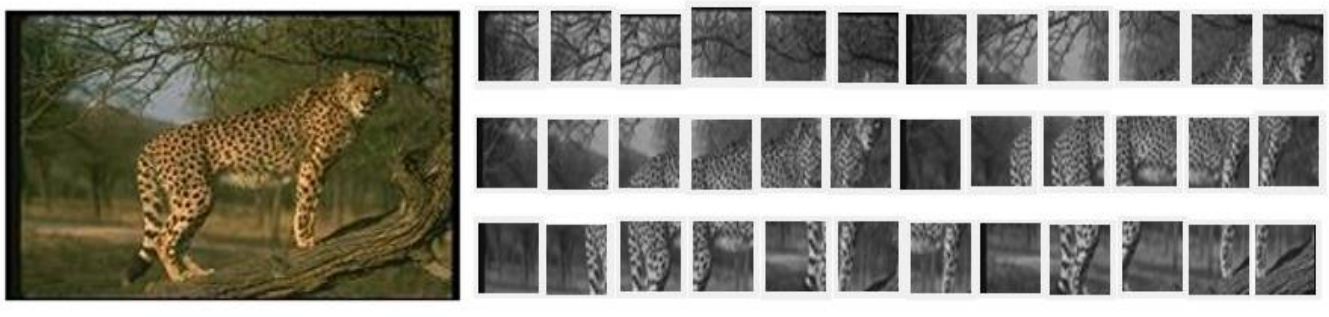

Fig. 2. *A sample image of a leopard chosen from the Caltech-101 Objects collection, along with 36 tiles obtained using overlapping spatial tiling*

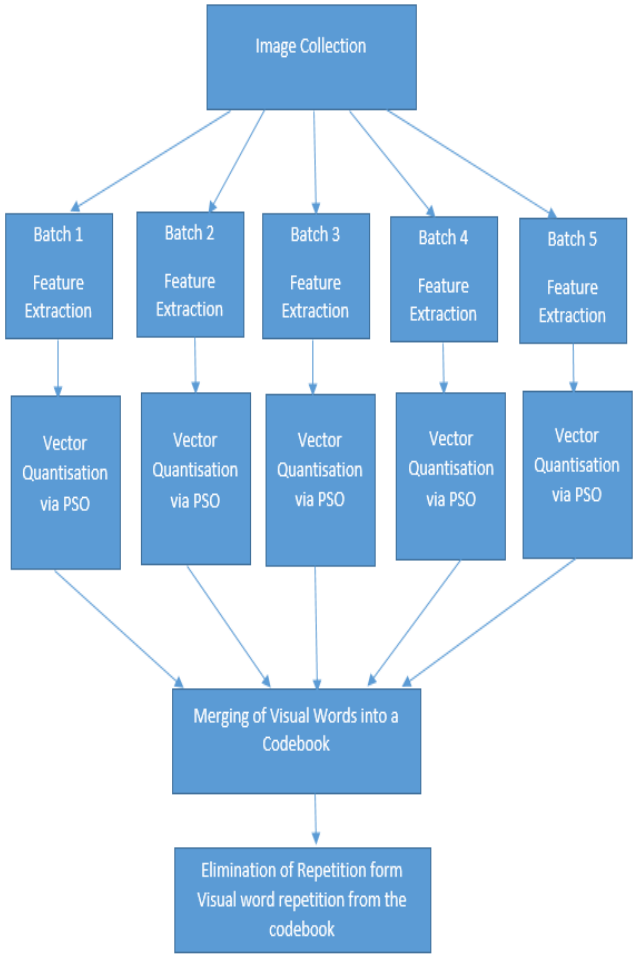

Fig. 3. *The block diagram illustrating steps of the proposed BOV Codebook development approach*

3.2. Batch vector quantisation using PSO

After features have been extracted from all the images in the collection to be classified using the 3-layered Stacked-Autoencoder, the image features need to be quantized into Visual Words using PSO. Although compared to SIFT and SURF, the image features generated for any given image collection with the Stacked-Autoencoder is considerably less, when the image collection is large, the number of image features generated using Stacked Autoencoder may still be numerous enough to cause lengthy computation during the implementation of the PSO clustering [59].

To ensure fast implementation of the PSO algorithm when applied to large number of image features (10,000 and above), the proposed BOVW codebook process groups the image features to be quantised into batches. The use of batch processing also allows the quantisation task to be divided among multiple computers. Figure 3 is the block diagram of the proposed codebook development framework.

3.2.1 Proposed Cluster Initialisation Algorithm: The primary goal of this clustering initialisation algorithm is to estimate the number of clusters within a set of image features based on the dimensionality of the image features and distribution of dimensional values. It also provides suitable representations of these clusters, which can then be fine-tuned by the PSO clustering process. The algorithm achieves these goals by dividing the image feature's multidimensional space into regions, identifying active regions. It then uses average pooling to generate representative samples from the active regions, while ignoring locations which do not attract any sample (dead centers).

By dividing the multidimensional space into regions, the initialisation algorithm assumes that the values of each dimension conform to the normal distribution, and a surplus number of random location are generated using the mean and standard deviation of the dimensional values, thereby minimising the likelihood of presenting closely similar initialisation points. The algorithm then statistically analyses the number of samples each of these points attracts to identify the active points. The implementation steps for the arbitrary image feature set $\boldsymbol{X}$ with N members is shown in Table 1.

Table 1. The proposed steps for the initialisation of PSO Clustering

Steps	ALGORITHM 1: PSO Clustering initialisation
I.	Calculate the mean m, and standard deviation d of each dimension in set X, and used them to generate a 5 membered set values as shown in Equation 4 for each dimension in X, where i is the index of the dimension. $P_i = \{m - 2d, m - d \;\; m, m + d, m + 2d\}$ (4)
II.	From each P_i, randomly constitute a column vector V_i, with length $0.2 * N$, and concatenate all the vectors to yield a matrix Y, whose rows represents locations in the multidimensional space.
III.	Evaluate the Euclidean distances between each row in matrix X, with all the rows in matrix Y.
IV.	Record the number of times each row in Y scores the minimum Euclidean distance with a row in X. entre the scores in a vector W.
V.	Calculate the means and standard deviation of the scores recorded in W.
VI.	Ignore any row in Y, whose score is less than mean minus standard deviation, and use average pooling to represent the rows in X attracted to the same row in Y (rows not being ignored).

While the locations identified in Step VI of this algorithm are good enough for use as the cluster centers, the locations will be improved when applied as the initialisation samples for the implementation of PSO.

3.2.2 Image Feature Clustering using PSO: The superiority of PSO clustering over K-Means clustering lies in its ability to track the movement of each particle, and pick the best location recorded at the end of the clustering process [74]. Therefore, this implementation of PSO clustering records the set of locations obtained at the end of each iteration and along with their respective measure of fitness.

The measures of the fitness of locations yielded at the completion of an iteration is the sum of the Euclidean distances between each sample in the set and the swarm particle it is attracted to during the iteration. Given that the set of swarm fitness recorded during the PSO clustering is $D=(d_1, d_2, d_3 \ldots\ldots\ldots d_n)$, where n is the number of iterations, the set of particles location with minimum fitness value will be chosen as the cluster centers as shown in Equation 5.

$$F = D_{min} \quad (5)$$

If the clustering process is completed in M iterations, each particle is expected to have gone through M locations in the multidimensional space. Given that the set $\overline{D}_t = (\overline{D}_1, \overline{D}_2, \overline{D}_3 \ldots \overline{D}_M)$ contains average distances the particle registered at each of its locations, the best location will be the location which records the minimum average Euclidean distance.

Although the use of this implementation of PSO clustering for vector quantisation can reduce thousands of image features into a few hundred visual words, the independence of each batch quantisation can result in the occurrence of the same Visual Word more than once in the final codebook when the visual words obtained from all the batches are merged into a single set. This problem is tackled using Visual word similarity analysis in Sub-Section III-C.

3.2.3 Visual word similarity analysis: The final codebook is initiated using any visual word from the merged set, after which other visual words are progressively added. A visual word from the merged set is added to the final codebook if and only if it does not record a Euclidean distance less than the threshold value with any visual word that is already in the final codebook. Therefore, to prevent repetition of visual words in the final codebook, a similarity threshold needs to be established via statistical analysis. This similarity criterion must be exceeded by any two visual words for both to exist in the same codebook. An experimental determination of the similarity criterion is demonstrated in Section 4.1.

4. Experiments

Using experiments on image collections constituted from Caltech-101 images as shown in Table 2, this section determines the appropriate statistical estimate for the similarity criterion, $E_{threshold}$ for BOVW codebook development using unsupervised machine learning via Self Organisng Map (SOM) implemented, then evaluates the performance of vector quantisation via PSO in comparison to other existing vector quantisation algorithms for BOVW codebook development. Finally, this section compares the performance of the unsupervised image categorisation built on the adaptive BOVW modelling with the performance of unsupervised Image categorisation via Hyper-graph partitioning. The experimental image collection.

Table 2. The Description of the 3 Experimental Image Collection

Collection	Object List
A	Airplane, Motorbike, Faces, Car
B	Airplane, Motorbike, Face, Car, Watch
C	Airplane, Motorbike, Faces, Car, Watch, Ketch

For the experimental determination of a similarity threshold ($E_{threshold}$) value for the proposed BOVW codebook development approach, this study adopts the same 3 image collections (described in Table 1) used by *Huang et al.* [12]. In this experiment, 100 images are chosen randomly from each category, and converted to grayscale. To improve the possibility of capturing objects during the codebook development, overlapping spatial tiling is employed, where the mask size of 0.25*L-by-0.25*B (L=Length of Image, and Breadth of Image), yielding 36 tiles from each image all which are resized into 40-by-40 pixels.

For each experimental categorization process, the 3-layered Stacked-Autoencoder is trained using all the spatial tiles obtained from the images in the entire set to be categorized, after which the trained Stacked-Autoencoder is then used to convert each tile in the experimental set to a 100-dimensioned vector. The resulting set is quantised into Visual Words using the proposed batch process with varying similarity threshold values. In this experiment, 5000 image features are handled in each batch during the vector quantisation process. It is common for clustering algorithms to perform hundreds of iterations before attaining convergence, especially when handling thousands of high dimensional data samples. Therefore, this implementation of PSO clustering is designed to exit the process after 50 of iterations to guarantee efficiency.

To boost categorisation accuracies, spatial incoherency is minimised during the image BOVW modelling using Level 2 spatial pyramid implementation [77]. Using PLSA, the dimension of the BOVW representations are reduced to 25 latents topics [78, 37], and the resulting set of image representations are clustered into the respective number of categories using SOM. After the clustering, each object is annotated based on the highest object category present in the cluster it belongs, and the accuracy of the process is evaluated by counting the number of annotations matching the ground truth.

4.1. Experimental determination of BOVW codebook visual word similarity criterion

The most important factor in the establishment of the similarity threshold for the merged visual words set, is the statistical distribution of the pairwise similarity distances. Where the pairwise Euclidean distances recorded by a set of visual words is represented by the set E in Equation 6, the mean E_{mean} can be calculated as shown in Equation 7, where x and y represent the position of the Euclidean distance on the proximity matrix holding all the possible pairwise Euclidean distances.

$$E = \{E_{1,1}, E_{1,2}, E_{1,3} \ldots\ldots\ldots E_{N,N}\} \quad (6)$$

$$E_{mean} = \frac{\sum_{x=1,y=1}^{N} E_{x,y}}{n(E)} \quad (7)$$

Figures 4A and 4B demonstrates the effects of varying the similarity threshold values between 0.25*Mean Euclidean distance to 2*Mean Euclidean distance on the number of visual words detected from merged visual words sets and the corresponding classification accuracies respectively.

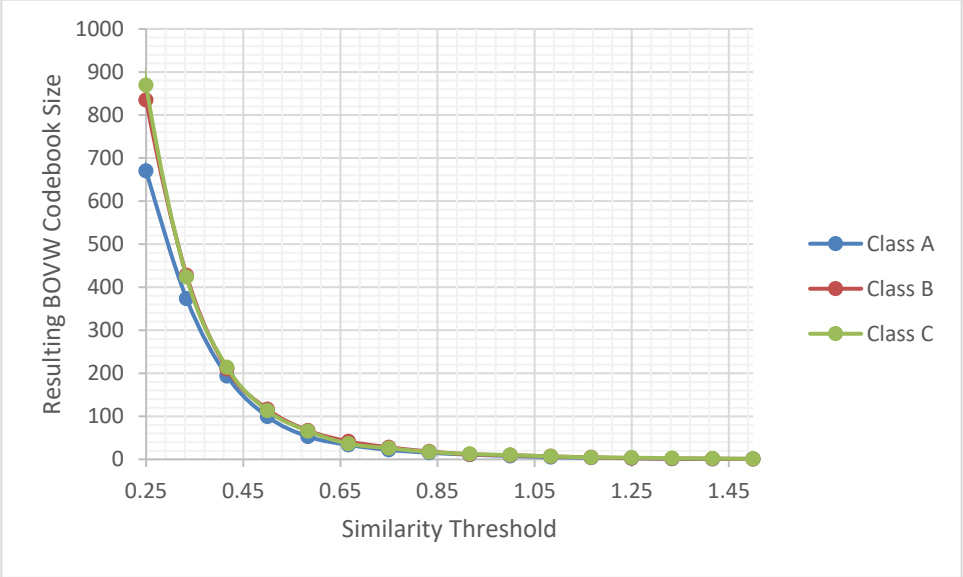

Fig. 4A. *The graphical representation of variation in the number of visual words detected in response to the changes in* $E_{threshold}$

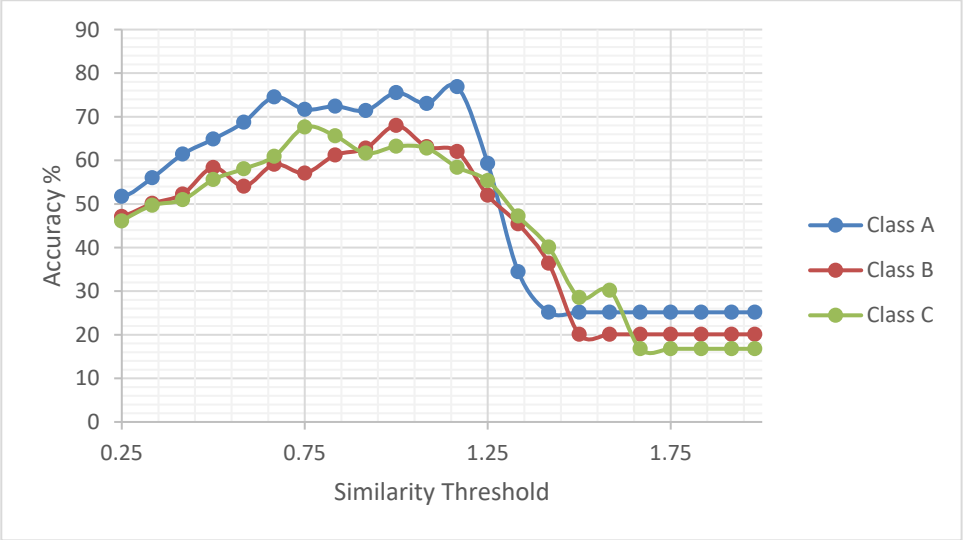

Fig. 4B. *The graphical representation of variation in Classification accuracy in response to the changes in* $E_{threshold}$

The graphs indicate that optimum classification accuracy is obtained for each image collection when $E_{threshold}$ is approximately equal to the Mean pairwise Euclidean distance of the merged set of visual words.

4.2. A comparison between the modified PSO and other BOVW vector quantisation techniques

This Sub-Section compares the performance of the proposed vector quantisation via modified PSO clustering with existing methods for the generation of BOVW codebook in the categorization of Collection. In this experiment, the proposed batch PSO vector quantisation is applied for codebook development using the visual word similarity criteria of E_{mean}. The implementation of the traditional K-means and the K-Means + HIK in this experiment both adopts the number of visual words in the Codebook developed by the batch PSO. While X-Means adopts the half the value of the PSO codebook size as its minimum and two times of the PSO codebook size as its maximum.

In all cases, a Stacked-Autoencoder with 100 neurons at its output is used as the image feature extraction algorithm. Table 3 is a summary of the performance recorded by the codebooks developed by these algorithms. Table 3 confirms the superiority of our proposed PSO based vector quantisation technique over the notable existing techniques. Due to the evaluation of clustering performance at the end of every iteration and subsequent comparison of the performances recorded at the end of the clustering process, the PSO based technique was able to identify much better set of centers unlike the K-means algorithm which limits its choice to the set of centers obtained at the end of the clustering process.

Table 3. A Comparison of The Proposed Batch PSO BOVW Codebook development With Other Methods

Vector Quantisation	Collection A	Collection B	Collection C
Modified PSO	89.84%	83.90%	83.43%
K-Means	80.68%	77.54%	82.15%
K-Means + HIK	85.31%	80.21%	81.84%
X-Means	87.44%	80.23%	83.87%

Unlike the traditional BOVW codebook (typically with 1000 visual words), the use of the proposed PSO based vector quantisation technique which yielded averages of 23, 25 and 35 visual words for Collection A, Collection B and Collection C respectively in this experiment ensures that heavy computation due to high number of visual words is avoided during the evaluation of the resulting BOVW image model. A comparison between the accuracies shown on Table 4 confirms that better accuracy can be achieved by using the higher quality visual words provided by the modified PSO.

While the modification of the K-Means algorithm by substituting Euclidean distance with HIK for vector similarity comparison yielded improvement in accuracies in Collection A and B, it has failed to record any improvement with Collection C when compared to the traditional K-Means. However, the proposed PSO based codebook development has been able to record leading performances across all three collections.

Although, the X-Means clustering's approach of varying the number of clusters and evaluating the clustering performance records better BOVW classification performance than K-Means and K-means + HIK, its lack of proper cluster initialisation method renders it classification to fall behind that of the proposed PSO based approach, and its approach of implementing clustering several times also renders the time taken to completion to be 20 times that of the proposed modified PSO, therefore it is less efficient than the proposed method.

4.3. A comparison between the unsupervised categorisation using adaptive BOVW/SOM and hypergraph partitioning

This section compares the performance of the proposed BOVW approach when combined with the Unsupervised Region of Interest detection proposed in [79], and compares the result to the unsupervised categorisation of images via Hypergraph partitioning propose by Huang *et al* [12]. The result obtained with the two approaches are presented in Table 4.

Figure 5 is a demonstration of the 4 Region of Interests (ROI) along with the sample Caltech image from which they were detected by using a mask with a dimension of 0.5*L by 0.5*B (L and B are the length and breadth of the image) to search the image space, with the goal of limiting the amount of background information captured during the BOVW image modelling and detecting visual words that frequently occur together in the image collection, thereby elimination spatial incoherency in the BOVW representation [79]. The ROIs detected from all the images in an experimental collection are then modelled using the adaptive BOVW technique proposed in this paper, with Emean as the similarity criteria ($E_{threshold}$). The result obtained is compared to the unsupervised image categorisation using hypergraph partitioning in Table 4.

As shown in Table 4, the BOVW based unsupervised categorization framework presented in successfully eliminated the spatial incoherency commonly associated with BOVW using unsupervised Region of interest detection, and Cross-Region Matching [80], thus provides a suitable means of demonstrating the benefit of BOVW. The Table confirms the superiority of the unsupervised image classification built on the adaptive BOVW and SOM clustering [79] over the unsupervised image classification via hyper-graph partitioning.

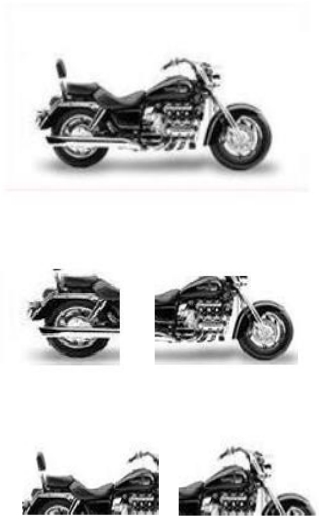

Fig. 5. *A demonstration of the content of four ROI detected from a sample image*

Table 4 A Comparison Between the accuracies obtained via Adaptive BOVW Modelling and Hypergraph Partitioning

Collection	Spatial Pyramid	Visual Sentence Modelling	Hypergraph Partitioning [3]
A	86.23%	99.66%	98.53%
B	83.42%	99.66%	97.38%
C	81.20%	99.21%	96.05%

5. Conclusion

This paper successfully demonstrates the application of Deep Feature Learning via Stacked-Autoencoder to the image feature extraction stage of BOVW modelling, where it enables the performance of the image classification framework to be optimised by varying the number of neurons employed at the different layers of the deep feature learning, resulting in a change in the dimensionality of the image feature vectors.

This paper also demonstrates the application of PSO clustering with a novel cluster initialisation technique in a batch vector quantisation process for the efficient development of BOVW codebook along. Perhaps, the greatest benefit of the approach is its scalability, which allows it to adjust its computation to be proportional to the number of images being categorised. In addition, the experimental results demonstrate that the misclassifications experienced because of over-fitting created by excessive number of visual words, can be removed through the application of the adaptive codebook development approach. The visual words obtained in using this approach also correlate to the objects or semantic contents in the image collection which makes the proposed approach an important step in the semantic content-based annotation of the images in the collection.

Furthermore, the adoption of a batch BOVW codebook development approach is an important step towards the implementation of Incremental Learning, since it yields a codebook whose visual words set can increase in quality and quantity and facilitate the application of parallel computation, thereby allowing the time required for the BOVW codebook to be significantly reduced

References

[1] H. Deljooi and S. J. Jassbi, "A Multi Criteria Decision Making Based Approach for Semantic Image Annotation," *International Journal of Computer-Aided Technologies,* vol. 2, no. 1, pp. 17-30, 2015.

[2] H. Deljooi and A. Eskandari, "A Novel Semantic Statistical Model for Automatic Image Annotation Using the Relationship between the Regions Based on Multi-Criteria Decision Making," *International Journal of Electrical & Computer Engineering,* vol. 4, no. 1, pp. 37-51, 2014.

[3] H. Sahlani and M. Hourali, "A Novel Semantic Statistical Model for Automatic Image Annotation Using Ontology," *Majlesi Journal of Multimedia Processing,* vol. 4, no. 2, pp. 1-10, 2015.

[4] P. Mookdarsanit and L. Mookdarsanit, "An Automatic Image Tagging of Thai Dance's Gestures," in *Joint Conference on ACTIS & NCOBA*, Pranakhon Si Ayutthaya, 2018.

[5] D. Zhang, M. M. Islam and G. Lu, "A Review on Image Annotation Techniques," *Pattern Recognition,* vol. 45, no. 1, pp. 345-362, January 2012.

[6] J. Cao, C. Wu, L. Chen, H. Cui and G. Feng, "An Improved Convolutional Neural Network Algorithm and Its Application in Multilabel Image Labeling," *Computational Intelligence and Neuroscience,* vol. 2019, pp. 1-12, 2019.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," Cornel University, 2015.

[8] A. Olaode and G. Naghdy, "Local Image Feature Extraction using Stacked-Autoencoder in the Bag-of-Visual Word modelling of Images," in *5th IEEE International Conference on Computer and Communication*, Chengdu, 2019.

[9] J. Deng, A. C. Berg, K. Li and L. Fei-Fei, "What Does Classifying More Than 10,000 Image Categories Tell Us?," in *European Conference on Computer Vision*, Crete, 2010.

[10] A. Olaode, G. Naghdy and C. Todd, "Unsupervised Classification of Images: A Review," *International Journal of Image Processing,* vol. 8, no. 5, pp. 325-342, 2014.

[11] G. Zazzaro and A. Martone, "ECF-means – EnsembleClusteringFuzzification Means," in *he Eighth International Conference on Advances in Information Mining and Management*, Barcelona, 2018.

[12] Y. Huang, Q. Liu, F. Lv, Y. Gong and D. N. Metaxas, "Unsupervised Image Categorization by Hypergraph Partition," *IEEE Transactions On Pattern Analysis And Machine Intelligence,* vol. 33, no. 6, June 2011.

[13] W. Zhang, X. Wang, D. Zhao and X. Tang, "Graph Degree Linkage: Agglomerative Clustering on a

Directed Graph," in *European Conference on Computer Vision*, Florence, 2012.

[14] G. Kim, C. Faloutsos and M. Hebert, "Unsupervised Modeling of Object Categories Using Link Analysis Techniques," in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, 2008.

[15] R. Balakrishnan and K. Kumar, "An Application of Genetic Algorithm with Iterative Chromosomes for Image Clustering Problems," *International Journal of Computer Science*, vol. 9, no. 1, pp. 60-67, 2012.

[16] R. Datta, D. Joshi, j. Li and J. Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," *ACM Computing Surveys*, vol. 40, no. No. 2,, pp. Article 5:1-60, Apri 2008.

[17] H. H. Wang, D. Mohamad and N. A. Ismail, "Semantic Gap in CBIR: Automatic Objects Spatial Relationships Semantic Extraction and Representation," *International Journal Of Image Processing (IJIP)*, vol. 4, no. 3, 2010.

[18] J. Xu, H. Li, P. Liu and L. Xiao, "A Novel Hyperspectral Image Clustering Method With Context-Aware Unsupervised Discriminative Extreme Learning Machine," *IEEE Access*, vol. 6, pp. 16176 - 16188, 2018.

[19] F. Baig, M. Rashid, M. A. Javid, A. Rehman, T. Saba and A. Adnan, "Boosting the Performance of the BoVW Model Using SURF–CoHOG-Based Sparse Features with Relevance Feedback for CBIR," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, vol. 1, no. 4, pp. 1-20, 2019.

[20] A. Faheema and R. Subrata, "Feature Selection using Bag-Of-Visual Words Representation," Centre for AI and Robotics (CAIR), Bangalore, 2010.

[21] P. Tirilly, V. Claveau and P. Gros, "Language Modelling for Bag-of-Visual Words Image Categorisation," CNRS-IRSA Image processing and computer vision, Rennes, 2008.

[22] J. Wu, W.-C. Tan and J. M. Rehg, "Efficient and Effective Visual Codebook Generation Using Additive Kernels," *Journal of Machine Learning Research*, vol. 12, pp. 3097-3118, 2011.

[23] J. Zhaoyin, C. Tsuham and Z. Yimeng, "Image Retrieval with Geometry-Preserving Visual Phrases," School of Electrical and Computer Engineering, Cornell University, 2010.

[24] K. Srinivas and V. Srikanth, "A Scientific Approach for Segmentation and Clustering Technique of Improved K-Means and Neural Networks," *International Journal of Advanced Research in Computer Science and Software Engineering,*, vol. II, no. 7, pp. 183-189, 2012.

[25] J. Guo, Z. Qiu and C. Gurrin, "Exploring the optimal visual vocabulary sizes for semantic concept detection," in *International Workshop on Content Based Multimedia Indexing (CBMI)*, Veszprem, 2013.

[26] R. Rane, B. K. Khadse and R. Suralkar S, "A review of Object Recognition Using Visual Codebook," *International Journal of Computer Science and Mobile Computing*, vol. II, no. 2, pp. 74-79, 2013.

[27] J. C. Van Gemert, C. G. Snoek, C. J. Veenman, A. W. Smeulders and J.-M. Geusebroek, "Comparing compact codebooks for visual categorization," *Computer Vision and Image Understanding*, no. 114, p. 450–462, 2010.

[28] Z. Suhail, A. Mahmood, E. Denton and R. Zwiggelaar, "Bag of visual words based approach for the classification of benign and malignant masses in mammograms using voting-based feature encoding," in *14th International Workshop on Breast Imaging (IWBI 2018)*, Atlanta, 2018.

[29] R. Wang, K. Ding, J. Yang and L. Xue, "A novel method for image classification based on bag of visual words," *Journal of Visual Communication and Image Representation*, vol. 40, pp. 24-33, 2016.

[30] D. Chanti and A. Caplier, "Improving Bag-of-Visual-Words Towards Effective Facial Expressive Image Classification," in *13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018)*, Madeira, 2018.

[31] R. Mandal, P. P. Roy, U. Pal and M. Blumenstein, "Bag-of-Visual-Words for Signature-Based Multi-Script Document Retrieval," Cornell University, 2018.

[32] M. Law, N. Thome and M. Cord, "Bag-of-Words Image Representation: Key Ideas and Further Insight," in *Fusion in Computer Vision, Advances in Computer Vision and Pattern Recognition*, Cham, Springer International Publishing Switzerland , 2014, pp. 29-52.

[33] N. Singhal, N. Singhal and V. Kalaichelvi, "Image classification using bag of visual words model with FAST and FREAK," in *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, 2017.

[34] J. Jiang, D. Wu and Z. Jiang, "A correlation-based bag of visual words for image classification," in *2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC)*, Chongqing, 2017.

[35] W. Li and Y. Dong, "Scene classification based on the bag-of-visual-words and Doc2Vec models for high-spatial resolution remote-sensing imagery," *Journal of Applied Remote Sensing*, vol. 13, no. 2, 2019.

[36] K. Xu, W. Yang, G. Liu and H. Sun, "Unsupervised Satellite Image Classification Using Markov Field Topic Model"," *IEEE Geoscience And Remote Sensing Letters*, vol. 10, no. 1, pp. 130-134, January 2013.

[37] A. Bosch, A. Zisserman and X. Munoz, "Scene Classification via PLSA," Computer Vision and Robotics Group, University of Girona, Girona, 2006.

[38] P. Olukanmi, F. Nelwamondo and T. Marwala, "k-means-lite: real time clustering for large datasets," in *IEEE 5th International Conference on Soft Computing and Machine Intelligence*, Nairobi, 2018.

[39] C.-F. Tsai, "Two Strategies for Bag-of-Visual Words Feature Extraction," in *7th International Congress on Advanced Applied Informatics (IIAI-AAI)*, Yonago, 2018.

[40] J. Wan, D. Wang, S. C. Hoi, P. Wu, J. Zhu, Y. Zhang and J. Li, "Deep Learning for Content-Based Image Retrieval: A Comprehensive Study," in *IEEE Conference on Computer Vision and Pattern Recognition Workshiops*, Orlando, 2015.

[41] M. N. Najafabadi, F. Villanustre, T. M. Khoshgoftaar and N. Seliya, "Deep learning applications and challenges in big data analytics," *Journal of Big Data,* vol. 2, no. 1, 2015.

[42] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 39, no. 4, pp. 664-676, 2016.

[43] Y. Bengio, A. Courville and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 8, pp. 1798 - 1828, 2014.

[44] Y. Wu and R. Razavi, "An Introduction to Deep Learning: Examining the Advantages of Hierarchical Learning," Predictive Analytics, Santa Barbara, 2015.

[45] J. Zhang, Z. He, J. Zhang and T. Dai, "Cograph Regularized Collective Nonnegative Matrix Factorization for Multilabel Image Annotation," *IEEE Access,* pp. 1-1, 2019.

[46] A. Krizhevsky, I. Sutskever and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems,* vol. 25, pp. 1106-1114, 2012.

[47] T. Patel, M. Kapadia and J. Maisuria, "A Review on Content based Image Retrieval," *International Journal of Computer Applications,* vol. 132, no. 13, pp. 22-25, 2015.

[48] S. Wang, Z. Ding and Y. Fu, "Feature Selection Guided Auto-Encoder," in *Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, San Francisco, 2017.

[49] W. Chu and D. Cai, "Stacked Similarity-Aware Autoencoders," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, Melbourne, 2017.

[50] G. E. Hinton and R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science,* vol. 313, pp. 504 - 507, 2006.

[51] Q. Xu, C. Zhang, L. Zhang and Y. Song, "The Learning Effect of Different Hidden Layers Stacked Autoencoder," in *8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, 2016.

[52] Z. Wu and Y. Junqing, "A multi-level descriptor using ultra-deep feature for image retrieval," *Journal of Multimedia Tools Application,* vol. 78, no. 318, pp. 1-18, 2019.

[53] W. Wu and D. Sun, "Multiple deep CNN for image annotation," in *Tenth International Conference on Graphics and Image Processing*, Chengdu, 2018.

[54] C.-C. HSu and C.-W. Lin, "Unsupervised convolutional neural networks for large-scale image clustering," in *IEEE International Conference on Image Processing*, Beijin, 2017.

[55] L.-Y. Gui, L. Gui, Y.-X. Wang, L.-P. Morency and J. M. Moura, "Factorized Convolutional Networks: Unsupervised Fine-Tuning for Image Clustering," in *IEEE Winter Conference on Applications of Computer Vision*, Lake Tahoe, 2018.

[56] F. Radenovic, G. Tolias and O. Chum, "CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples," in *European Conference on Computer Vision*, Amsterdam, 2016.

[57] H. Bay, T. Tuytelaars and L. V. Gool, "SURF: Speeded Up Robust Features," ETH Zurich, Zurich, 2005.

[58] T. S. Shinde and A. K. Tiwari, "Pruning SIFT & SURF for Efficient Clustering of Near-duplicate Images," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, 2019.

[59] M. Balayil, G. S. Kumar and V. M. Anees, "Automatic Multilabelling of Images and Semantic Relation Extraction," in *Intelligent Signal Processing Conference (ICISP)*, Cherbourg, 2018.

[60] Y. Li, Y. Xu, J. Wang, Z. Miao and Z. Yafei, "MS-RMAC: Multiscale Regional Maximum Activation of Convolutions for Image Retrieval," *IEEE Signal Processing Letter,* vol. 24, no. 5, pp. 609 - 613, 2017.

[61] M. EL Agha and W. Ashour, "Efficient and Fast Initialisation Algorithm for K-means Clustering," *International Journal of Intelligent Systems and Applications,* vol. I, pp. 21-31, 2012.

[62] C.-F. Tsai, "Bag-Of-Words Representation in Image Annotation: A Review," *ISRN Artificial Intelligence,* vol. 2012, pp. 1-19, 2012.

[63] F. Jurie and B. Triggs, "Creating Efficient Codebooks for Visual Recognition," in *Tenth IEEE International Conference on Computer Vision*, Beijin, 2005.

[64] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning-Data Mining, Inference and Prediction, 2nd Edition ed., vol. II, Stanford: Springer, 2008, pp. 465-576.

[65] S. Salvador and P. Chan, "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms," in *16th IEEE International Conference on Tools with Artificial Intelligence*, Florida, 2004.

[66] M. Yan, "Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion," Virginia Polytechnic Institute and State University, Blacksburg, 2005.

[67] Y. Jun, N. Chong-Wah, G. H. Alexander and J. Yu-Gang, "Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study," City University of Hong-Kong, Hong-Kong, 2008.

[68] S. Battiato, G. M. Farinella, T. Meccio, G. Puglisi, D. Ravi and R. Rizzo, "Bags of Phrases with Codebooks Alignment for Near Duplicate Image Detection," in *Multimedia in Forensics, Security and Intelligence*, Firenze, 2010.

[69] A. Olaode, G. Naghdy and C. Todd, "Bag-of-Visual Words Codebook Development for the Semantic Content Based Annotation of Images," in *Signal Image*

Technology and Internet Based System, Bangkok, 2015.

[70] A. A. Olaode, G. Naghdy and C. A. Todd, "Efficient Region Of Interest Detection using Blind Image Division," in *Signal Processing Symposium* , Debe, Poland, 2015.

[71] D. Pelleg and A. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," in *17th International Conference on Machine Learning*, Stanford, 2000.

[72] K. Kersorn, S. Chimlek, S. Poslad and P. Piamsa-nga, "Visual content representation using semantically similar visual words," *Expert Systems with Applications*, vol. 38, pp. 11472-11481, 2011.

[73] D. Tsou and C. MacNish, "Adaptive Particle Swarm Optimisation for High-Dimensional Highly Convex Search Spaces," in *The 2003 Congress on Evolutionary Computation*, Canberra, 2003.

[74] A. L. Ballardini, "A tutorial on Particle Swarm Optimization Clustering," Cornell University, New York, 2016.

[75] N. Kamel, I. Ouchen and K. Baali, "A Sampling-PSO-K-means Algorithm for Document Clustering," in *Seventh International Conference on Genetic and Evolutionary Computing (ICGEC)*, Prague, 2013.

[76] C. Gong, H. Chen, W. He and Z. Zhang, "Improved multi-objective clustering algorithm using particle swarm optimization," *PL0S One*, vol. 12, no. 12, p. e0188815, 5 December 2017.

[77] S. Lazebnik, C. Schmid and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference*, Illinois, 2006.

[78] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, vol. 42, pp. 177-196, 2001.

[79] A. Olaode and G. Naghdy, "Elimination of Spatial Incoherency in Bag-of-Visual Words Image Representation Using Visual Sentence Modelling," in *International Conference on Image and Vision Computing New Zealand (IVCNZ)*, Auckland, 2018.

[80] Z. Gao, L. Wang and L. Zhou, "A Probabilistic Approach to Cross-Region Matching-Based Image Retrieval," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1191-1204, 2019.