Unsupervised Classification of Images: A Review

Abass Olaode *abass.olaode808@uowmail.edu.au*
School of Electrical Computer Telecommunication Engineering
University of Wollongong
Wollongong, 2500, Australia

Golshah Naghdy *golshah@uow.edu.au*
School of Electrical Computer Telecommunication Engineering
University of Wollongong
Wollongong, 2500, Australia

Catherine Todd *CatherineTodd@uowdubai.ac.ae*
School of Electrical Computer Telecommunication Engineering
University of Wollongong
Dubai, UAE

Abstract

Unsupervised image classification is the process by which each image in a dataset is identified to be a member of one of the inherent categories present in the image collection without the use of labelled training samples. Unsupervised categorisation of images relies on unsupervised machine learning algorithms for its implementation. This paper identifies clustering algorithms and dimension reduction algorithms as the two main classes of unsupervised machine learning algorithms needed in unsupervised image categorisation, and then reviews how these algorithms are used in some notable implementation of unsupervised image classification algorithms.

Keywords: Image Retrieval, Image Categorisation, Unsupervised Learning, Clustering, Dimension Reduction.

1. INTRODUCTION

The advent of computers and the information age has created challenges in the storage, organisation and searching of complex data especially when the quantity is massive. Therefore, it is not surprising that application of pattern recognition techniques has been found useful in image retrieval, where it has been helpful in managing image repositories. Pattern recognition enables the learning of important patterns and trends, which can be used in the indexing of the images in a repository. The applicable learning approaches can be roughly categorised as either supervised or unsupervised [1].The supervised classification of images based on patterns learnt from a set of training images has often been treated as a pre-processing step for speeding-up image retrieval in large databases and improving accuracy, or for performing automatic image annotation [2]. The block diagram of a typical supervised image classification process is shown in Figure 1. This training data is manually selected and annotated, which is expensive to obtain and may introduce bias information into the training stage [3].

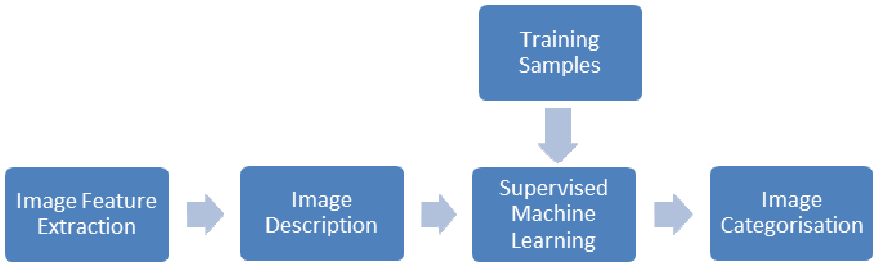

FIGURE 1: The Block diagram of a typical supervised Image categorisation process.

Alternatively, unsupervised learning approach can be applied in mining image similarities directly from the image collection, hence can identify inherent image categories naturally from the image set [3].The block diagram of a typical unsupervised classification process is shown in Figure 2. Due to the ability to support classification without the need for training samples, unsupervised classifications are a natural fit when handling large, unstructured image repositories such as the Web [2], therefore, this study considers the unsupervised classification of images an important step in the semantic labelling of the images in a large and unstructured collection because it enables grouping of images based on semantic content for the purpose of mass semantic annotation.

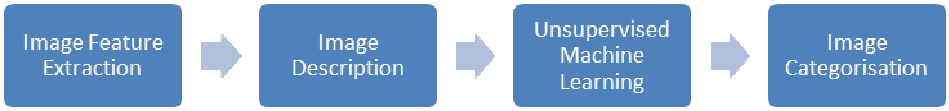

FIGURE 2: The Block diagram of an unsupervised Image categorisation process.

Due to the ability of unsupervised image categorisation to support classification without the use of training samples, it has been identified as a means of improving visualisation and retrieval efficiency in image retrieval [2]. It has also been identified as a means of matching low-level features to high-level semantics especially in learning based applications [3]. These qualities makes unsupervised image categorisation a likely solution for bridging the semantic gap in image retrieval [4]. In view of these, this study provides an overview of recently developed unsupervised image classification frameworks.

In the remainder of this paper, Section 2 provides a review of image modelling process, and analyses some image modelling approaches with some recent attempts at their improvement by the research community. Section 3 provides an insight into unsupervised learning algorithms, while Section 4 examines some notable implementations of unsupervised image classification. Finally, Section 5 suggest the future application of unsupervised image classification in the automated image annotation for image retrieval purposes, while Section 6 concludes the paper with a brief overview of the reviewed implementations with focus on their suitability in the semantic labelling of images.

2. IMAGE MODELLING

Features of a digital image such as colour, texture, shapes and the locations of these features on the image represent characteristics that enable the image to be distinguished from other images. As indicated by Figure 1 and Figure2, the first step in any supervised or unsupervised image classification is the detection and extraction features present in the image, which is then followed

by the development of a description that is based on the features extracted from each image. This section examines popular algorithms for achieving these two important image modelling functions during unsupervised image categorisation.

2.1. Image Feature Extraction

For reliable recognition, it is important that the features extracted from images be detectable even under changes in image scale, noise and illumination. To satisfy this need, keypoints corresponding to high-contrast locations such as object edges and corners on the image are typically regarded as good descriptive image features are sought. The features at these keypoints are typically described using image feature descriptors. Most popular image feature extraction algorithms consist of image feature detection and description components [5, 6]. Some popular image feature extraction algorithms mentioned in recent literature are discussed below.

The Shift Invariance Feature Transform (SIFT) is an image feature extraction algorithm that ensures the detection of keypoints that are stable and scale invariant [5]. In the SIFT algorithm, the keypoints are detected via a DoG (Difference of Gaussian) pyramid created using a Gaussian filtered copy of the image. Each of the detected keypoint is then described by a 128-dimensioned histogram of the gradients and orientations of pixels within a box centred on the keypoint.

Although SIFT remains one of the best descriptors in terms of accuracy, the 128-dimensions of the descriptor vector makes its feature extraction process relatively computationally expensive [7]. Khan et al. [8] explain that compared to the standard SIFT, a smaller size descriptor uses less memory and results in faster image classification. The authors [8] propose generating 96D-SIFT, 64D-SIFT and 32D-SIFT by skipping some orientation values during computation of the descriptors. Classification experiments on images from the Caltech dataset revealed that the 32D variant achieved 93% accuracy, the 64D and 96D versions recorded 95%, and the 128D achieved 97% accuracy [8]. The study reported that 128D, 96D, 64D and 32D recorded 59, 33, 18 and 11 seconds respectively to complete the classification task [8], which confirms that reducing the dimension of the descriptors reduces the amount of computation required for classification, thereby improving the speed of the process but leads to reduced accuracy. A similar result was also obtain by Ke and Sukthankar [9] when the dimension of the SIFT descriptor is reduced using Principal Component Analysis (PCA).

Rather than using DoG and image pyramid for the detection of keypoints as in SIFT, Speeded-Up Robust Features(SURF) uses the Hessian matrix in which the convolution of Gaussian second order partial derivatives with a desired image are replaced with box filters applied over image integrals (sum of grayscale pixel values), thereby reducing computation time [6]. To develop feature descriptions for the detected keypoints, each of keypoint is assigned a reproducible orientation using Haar wavelet responses in x and y directions for a set of pixels within a radius of 6σ, where σ refers to the detected keypoint scale [8]. The SURF descriptor is then computed by placing a 4 x 4 square window centre on the keypoint to produce a normalised 64D descriptor vectors [8]. The use of integral image representation at the keypoint detection stage of SURF ensures that the computational cost of applying box filter is independent of the size of the filter. This allows SURF to achieve much faster keypoint detection than SIFT by keeping the image size the same while varying only the filter size [8].

Although, SURF's performance is mostly similar to SIFT, it is unstable to rotation and illumination changes [10]. Liu et al. [11] noted that although SURF is capable of representing most image patterns, it is not equipped to handle more complicated ones. However, Khan et al [29] implemented classification experiments on images from David Nister, Indoor, Hogween and Caltech datasets to yield results that confirms that SURF's performance is as good as that of SIFT, with both recording 97% accuracies on Caltech dataset. The study however indicates that SURF's image matching time is higher at 80s compared to SIFT's 59s. Therefore this research finds SURF adequate enough to be considered for the purpose of feature extraction during image classification.

SIFT and SURF feature extraction algorithms can be regarded as sparse feature extraction algorithms because they only detect and describe features at chosen locations on an image. Extracting features from locations covering the entire area of an image rather than few selected location provides additional information which may improve the accuracy of image retrieval result [12]. Dalal and Triggs [13] proposed the Histogram of Oriented gradients (HOG) also known as Dense-SIFT which extracts and describes local image features from each of the uniformly spaced cells placed on an image.

A HOG description is developed for a cell by counting the occurrence of gradient orientations for the pixels within the cell in a manner similar to SIFT[14] The algorithm achieves more accurate description than SIFT by using overlapping local contrast normalization to make the result less variant to changes in illumination and shadowing [14]. This is achieved by calculating a measure of the intensity across a larger region of the image, called a block, and then using the value obtained to normalize all cells within the block [14]. Since the HOG descriptor operates on localised cells, it is invariant to geometric transformation of the image [14]. HOG was originally designed for the problem of pedestrian detection in static images but the use has been extended to other applications such as scene and object classification in static imagery [14, 13].
Due to recent increase in the use of computer vision application on mobile phones and other low power devices, research efforts are heading in the direction of development of feature extraction algorithms with minimum computation requirement and low power consumption. Examples of such algorithm include Oriented-FAST and Rotation-Aware BRIEF (ORB) [15] and Fast Retina Keypoint (FREAK) [16].

2.2. Image Description

After the extraction of features present in an image, there is a need for mathematical description of the image before supervised or unsupervised classification can be possible. Thus, the performance of image annotation is dependent on the reliability of the image feature representation (image mathematical model) [17].The most common approaches discussed in recent literatures use a normalised histogram or a vector to represent the number of times quantised features occurs on an image. The most popular of these methods is the Bag-of-Visual words (BOV) image model.

The BOV model is a popular image representation for classification purposes, which uses a visual-words histogram to effectively represent an image for image annotations and retrieval tasks [18, 19]. An important stage during BOV representation of images is the visual codebook development; a process that requires the use of K-means clustering to quantise the vectors representing image features into visual-words [17, 19, 20]. The computational requirement of this stage is very high and is therefore regarded as the most expensive part of the BOV modelling process, and attempts at reducing the computation time often lead to noisy visual-words [19].This study considers the limiting of visual-words to the unique vectors available in the set of image features extracted as a likely solution to this problem.

Currently, there are no proven methods for determining the number of visual-words to quantise image feature vectors into, during codebook development. Tsai [17] explained that although most implementations of the BOV modelling are based on 1000 visual-words, the number of visual-words is dependent on the dataset. Bosch et al. [12] used an arbitrary value of 1500 as the number of visual-words developed from SIFT features vectors of sample images in all experimentation involving BOV, while Verbeek and Triggs [21] quantise the SIFT descriptors used in their work into 1000 bins. The use of these approaches exposes the classification process to limited distinctiveness due to a small number of visual-words in the codebook, and high processing overhead when a codebook with too many visual-words is used [18]. Therefore, a research into the determination of the number of visual-words needed during BOV modelling will provide a means of eliminating some unnecessary computation overhead.

Another problem with BOV modelling of images is the loss of classification accuracy due to the disregard for the spatial location of the visual words during the modelling process [21, 20, 22].

Verbeek et al. [21] used Random Field theory to provide spatial information along with the BOV models for Probabilistic Latent Semantic Analysis (PLSA) classification of image regions, the same approach was also adopted by Xu et al. [38] for image classification via Latent Dirichlet Allocation (LDA). Zhang et al. [20] proposed the Geometry-preserving visual phrase that encodes more spatial information into the BOV models at the searching step of a retrieval system, thereby representing local and long-range spatial interactions between the visual words. However, this approach only seeks to improve search results by providing additional information to the image BOV models but does not improve the BOV modelling therefore may not be suitable for semantic-based image retrieval purposes. Lazebnik et al. [22] proposed the Spatial Pyramid in which histograms are computed for multi-level regions of an image, and then concatenated to form a single spatial histogram. This method achieved 64.6% during the classification of images from Caltech-101 [22].While this study considers the work of Lazebnik et al. [22] as an intuitive attempt at solving the spatial coherency problem in BOV modelling especially for semantic purposes; it however suggests that the classification accuracy needs to be improved further.

Bosch et al. [23] extend the concept of spatial pyramid to the development of an image signature known as Pyramid Histogram of Oriented Gradient (PHOG). This image representation combines pyramid representation with HOG and has been found to be effective in object classification [24], facial emotion recognition [25], and facial component based bag of words [26]. How some of the techniques discussed in this Section have been used in recent works on unsupervised image classification is examined in Section 4.

3. UNSUPERVISED LEARNING

This section reviews existing literatures in which unsupervised learning approaches have been successfully applied to image categorisation. In general, unsupervised learning attempts to base grouping decisions directly on the properties of the given set of samples without the use of training samples. Although Datta et al. [2] identified three main categories of unsupervised classification algorithms: clustering based on overall minimisation of objective function, pairwise distance clustering, and statistical modelling; such classification has limited the scope of unsupervised learning to clustering algorithms. Therefore, based on the extensive list of unsupervised learning algorithms provided by Hastie et al. [1], this paper recognises Dimension reduction algorithms and clustering algorithms as the two main unsupervised machine learning algorithms needed in unsupervised image categorisation. The recognition of these groups of algorithms provides a reasonable lead way into the diverse world of the application of unsupervised machine learning to image classification.

3.1. Dimension Reduction Algorithms

It is often necessary to reduce the dimension of samples in a dataset before the patterns can be recognised. For example, the application of BOV modelling may produce 1000 dimensioned image representations which can make the categorisation of a set image BOV representation computational inefficient especially when handling a large collection (1000 samples and above). This challenge can be minimised by estimating a rather crude global models that characterise the samples in the collection using descriptive statistics [1]. These descriptive statistics reduces the data dimensions by using a new set of variables based on properties observed at regions of high probability density on the sample space. Common unsupervised learning methods by which a descriptive statistics can be obtained include Principal Component Analysis (PCA), Non-negative matrix factorisation, and Independent component analysis (ICA) [1]. These are linear approaches which are not desirable as the means of achieving dimension reduction of images because image data possesses complicated structures which may not be conveniently represented in a linear subspace [27], therefore mapping them to a linear low dimensioned features space may incur significant loss of categorisation accuracy.

Several methods have been recently proposed for nonlinear dimension reduction. These methods are all based on the idea that the data lie close to an intrinsically low-dimensional nonlinear feature space embedded in a high-dimensional space [1]. Scholkopf et al. [28] proposed Kernel PCA as a mean of achieving non-linear dimension reduction [28]. Using non-linear functions,

Kernel PCA generates a kernel matrix for a given dataset, and then identifies a chosen number column on the kernel matrix with the largest eigenvalues [1]. Other popular non-linear dimension reduction methods include Isometric Feature Mapping (ISOMAP), Local Linear Embedding, and Local Multi-Dimensional Scaling (Local MDS) [1]. In general, these methods produce low-dimensional model of each sample in a collection by describing the sample in the terms of its approximate distances from a chosen number of nearest neighbours [1].

3.2. Clustering Algorithms

Clustering algorithms groups the samples of a set such that two samples in the same cluster are more similar to one another than two samples from different clusters, Clustering methods can be categorised into two broad classes: non-parametric and parametric methods. Non-parametric clustering involves finding natural groupings (clusters) in a dataset using an assessment of the degree of difference (such as Euclidean distance) between the samples of the dataset. It requires the defining a measure of (dis)similarity between samples, defining a criterion function for clustering, and defining an algorithm to minimise (or maximise) the criterion function. Popular non-parametric clustering algorithms include k-means, Hierarchical clustering algorithms and Spectral clustering.

3.2.1. Non-Parametric Clustering

K-mean clustering is the most widely used nonparametric technique for data clustering [2]. It represents each category in a given dataset with a centre obtained after repeated optimisation of an overall measure of cluster quality known as the objective function. The result of K-means clustering algorithm is sensitive to the initial centres used in the clustering process [29]. For example, randomly picking the initial centres may lead to accidentally picking too many centres that attracts few or no members while most of the samples allocated to a few of the centres. El Agha and Ashour [29] demonstrated that classification results can be improved when the overall shape of the dataset is considered during the initialisation phase of the K-means algorithm.

The K-means algorithm is very similar to the unsupervised Artificial Neural Network Algorithm known as Self Organising Map (SOM). The ability of all ANNs to process complex or high dimensional data at high speed makes SOM desirable for image classification [30]. In a SOM, the hidden layer consists of a matrix or a vector of neurons arranged in a grid, hexagonal or random pattern. In response to an input pattern, the neurons compete to be activated and the neuron whose weight has the smallest Euclidean distance from the input pattern is selected. The network updates the weight of the chosen neuron and its neighbours using Kohenen learning rule pattern and re-arranged its topology such that it correlates with the input vector space, thereby ensuring the same neuron will be chosen in response to subsequent input pattern similar to the current input [31]. Hastie et al. [1] considers SOM to be a constrained version of K-means algorithm in which the performance depends on the learning rate and the distance threshold, and stated under the same condition, SOM will outperform K-means clustering, if the constraints are adequate [1]. Therefore the determination of these constraints and the number of clusters are important for maximum accuracy to be achieved when using SOM for data clustering. Decision trees and Associative rules also provide simple rules by which each samples of an image dataset can be labelled.

Although the K-mean algorithm is very effective when the centres are positioned to capture the distribution of the category [1] and for the process to be credible, the number of clusters specified at the beginning of the process must closely match the number of categories present in the dataset. In contrast, hierarchical clustering methods do not require such specifications, but requires the user to specify a measure of dissimilarity between (disjoint) groups of observations. Hierarchical clustering creates a nested sequence of partitions in which the entire dataset is considered to be a single, all inclusive cluster at the highest level of the hierarchy, while each cluster at the lowest level contains a single sample. Hierarchical clustering can be implemented using either Agglomerative or Divisive approach in grouping samples into clusters [1].

In the Agglomerative approach, the clustering process starts at the lowest level and proceeds to the top, merging any two clusters whose members are considered to be similar. In the Divisive approach, the process starts from the all-inclusive clusters and repeatedly splits the dataset into smaller groups until the process attains a level where the members of each cluster are considered to be different from any other [1]. Hierarchical clustering based on the Agglomerative approach determines the affinity between samples using either single linkage, complete linkage or average linkage [1, 32]. Zhang et al. [32] explained that the Agglomerative approach is susceptible to noise and outliers because in calculating the link between two clusters to be merged, it does not consider the global similarities of the entire dataset, therefore it is not adequate for high-dimensional data such as images [32].

The use of K-means and hierarchical clustering in the unsupervised learning from an image dataset is often faced with the high dimensionality of the samples. Spectral clustering is a popular non-parametric [33] algorithm that achieves clustering through a combination of non-linear dimension reduction and K-means clustering, therefore it is preferred when the clusters are non-convex [1]. It achieves non-linear dimension reduction through the use of an undirected similarity graph to represent the pairwise distances between each sample and every other sample in the dataset, from which the normalised eigenvectors of each dimension is obtained and the desired number of columns is chosen based on their eigenvalues. The dataset samples represented by the normalised eigenvectors are then clustered using the k-means algorithm [12].

3.2.2. Parametric Clustering Methods

While on-parametric methods infer the underlying pattern structure from the dataset, parametric approaches impose a structure on the dataset. Parametric learning assumes the samples of the dataset can be represented by a probability function made up of several components [1].In parametric clustering methods, each sample in a set is described as a combination of a finite number of functions and samples with similar combinations as assumed to be in the same cluster. The use of probabilistic parametric clustering method such as Gaussian Mixture Model (GMM) [2] and Topic-based model [34] has been shown to be successful in a wide variety of applications concerning the analysis of continuous and discrete data, respectively [33].

Given a dataset, GMM fits a single probability density function to the entire set [35]. This function is assumed to be a mixture of a finite number of Gaussian functions as shown by Equation 1 and Equation 2[36]:

$$f(X, \varnothing) = \sum_{k=1}^{k} p_k g(X;\ m_k, \sigma_k)$$

.....1

Where

$$g(X;\ m_k, \sigma_k) = \frac{1}{(\sqrt{2\pi}\sigma_k)^D} e^{-\frac{1}{2}\left(\frac{\|X-m_k\|}{\sigma_k}\right)^2}$$

.....2

In Equation 1, Pk is the mixing probability for the Gaussian density function k in the mixture, while mk and σ_k are its mean and standard deviation respectively. These parameters are estimated through model fitting using Expectation-Maximisation (EM) process [36].

In GMM, knowledge of the probability density function parameters for a dataset enables the representation of each of its samples with a vector whose dimension is the same as the number of Gaussians in the mixture. While K-means clustering is regarded as hard clustering model because it exclusively maps each sample to a cluster, the GMM is considered a soft clustering method because it does not exclusively place a sample into any of the available clusters but describes the probability of its membership of each of the clusters.

Since each data sample is represented with a vector at the end of a GMM process, it is possible to represent these vectors in a multi-dimensional Euclidean space. Liu et al. [35] explained that this representation may reveal naturally occurring data patterns on, or close to subgroups within the data set and proposed the Locally Consistent Gaussian Mixture Model (LCGMM) which exploit these patterns to improve the learning performance of GMM. Experimentation conducted by the authors on Yale face and Breast cancer datasets revealed accuracies of 54.3 % and 95.5% respectively which is better than the 29.1% and 94.7% recorded by the conventional GMM [35].

Topic-based models such as PLSA and LDA are soft clustering techniques that are similar to GMM. Hoffman [34] presented the PLSA (also known as the Aspect model) for categorising collections of textual documents. Given D = d_1..... d_N is a set of BOV representations of images and a corresponding W = w_1 ..... w_V, a set of visual vocabularies. In the PLSA modelling a joint probability model over D X W with a set of unobserved variables Z = z_1 ..... z_k is defined by the mixture in Equation 3 and Equation 4[34, 12]:

$$P(d,w) = P(d)P(w|d)$$

…….3

Where

$$P(w|d) = \sum_{z \in Z} P(w|z)P(z|d)$$

……. 4

P(w|z) and P(z|d) are the topic specific distributions for the entire set and topic mixtures for each image respectively. The model is parameterised as shown in Equation 5[34].

$$P(d,w) = \sum_{z \in Z} P(z)P(d|z)P(w|z)$$

……. 5

Similar to GMM, the model parameters are estimated using the EM algorithm, at the end of which each image in the dataset is represented by the topic mixture P(z|d) and images with similar topic mixtures are considered to belong to the same cluster. The advantage of PLSA lies in its use of generative models obtained from the BOV representation of images for image modelling, rather than directly using the BOV representations; a step which enables the discovery of latent topics from the image data [37]. Since a group of related words is mapped to one latent topic at the end of Topic-based modelling, the resulting image representation has a reduced dimension compared to the BOV representation [34, 38].

Blei et al. [37] noted that PLSA does not provide a proper probabilistic model at the document level because the number of latent topics used to model each document grows linearly with the

size of the dataset which may lead to over-fitting. The authors therefore proposed Latent Dirichlet Analysis (LDA) which provides additional regularisation by encouraging the topic mixtures to be sparse [37]. Verbeek et al. [21] noted that LDA only outperforms PLSA when classifying small number of documents with many topics, therefore, considered PLSA to be computationally more efficient than LDA [21].

In general, Parametric models are advantageous since they provide principled ways to address issues like the number of clusters, missing feature values etc.[33]. They also combine dimension reduction capability with soft clustering [34, 12, 37, 38]. The unfortunate drawback is that they are only effective when the underlying distribution of the data is either known, or can be closely approximated by the distribution assumed by the model [33]. However similar shortcomings can be attributed to squared error based clustering algorithms such as K-means[33].

4. COMPARISON OF RELATED WORK

Unsupervised image classification is useful in the annotation of images in a large repository. In such a scenario, it can enable images to be grouped into a manageable number of clusters such that semantic labelling can be applied conveniently and efficiently. This section review literatures in which unsupervised machine learning have been applied to image categorisation.

Work	Year	Feature extraction	Image Model	Unsupervised learning approach
Xu et al. [23]	2013	SIFT	BOV	LDA / Markov Random Field / Bayesian Information criterion (BIC)
Zhang et al. [33]	2012	LBP, pixel intensity	Image Feature Histograms	Hierarchical clustering (GDL-U, AGDL)
Huang et al. [3]	2011	Dense SURF / PHOG	KNN/Hyper-graph	Spectral clustering
Mole and Ganesan [39]	2010	Local Binary Pattern	Texture histogram	K-Means
Duong et al. [40]	2008	HSV and Canny Edge Orientation histogram	Hierarchical tree	Tree matching
Kim et al. [41]	2007	Harris-Affine detector / SIFT descriptor	Page-Rank/Visual Similarity Network	Spectral clustering
Bosch et al. [12]	2006	SIFT / HOG	BOV	PLSA / KNN
Graum and Darrell [42]	2006	SIFT	Multi-resolution histogram/Similarity Graph	Spectral clustering
Todorovic and Ahuja [43]	2006	Pixel gray levels	Multiscale segmentation tree	Decision trees

Lee and Lewicki [44]	2002	Image texture	Pixel patches	ICA
Le Saux and Boujemaa [45]	2002	Pixel gray level, Fourier power spectrum and edge orientation	Feature histogram	PCA / Adaptive Robust Competition Clustering

TABLE 1: A comparison of some notable implementations of unsupervised image categorisation.

4.1. Methodology of Image Modelling

From Table 1, it can be observed that SIFT and SIFT-based algorithms are the most popular image feature extraction algorithm for the implementation of unsupervised image classification, while images are mostly described in terms of feature histograms. Bosch et al. [12]shows that dense descriptors (HOG) outperform the sparse ones, especially in the classification of images of scenes, where they enable the capturing more information than sparse features, and produce improved classification accuracy when image colour information is captured during feature extraction. This superiority was confirmed in the supervised classification by Verbeek et al. [21] the average accuracy of 61.3 % recorded by HOG descriptors was improved to 75.2 % by combining HOG with colour descriptors.

Table 1 also reveals that unsupervised image classification is yet to exploit the advantage of BOV modelling, especially it's potential in supporting Semantic-Based image retrieval. The work of Huang et al. [3] is of particular interest in this study not only because of its use of the combination of PHOG and Dense SURF features to develop representation for each image, but mainly because of its the use of Regions of Interest (ROI) of each image rather than the entire image, an approach with the potential to capture spatial information when applied along with BOV modelling.

4.2. Unsupervised Learning

As mentioned in the last section, the use of non-parametric clustering such as K-means, Hierarchical or SOM on high dimensional data samples such as histograms representing images is often a computationally inefficient process. This is perhaps the reason why none of these methods is popular as the unsupervised learning approach on Table 1. In Le Saux and Boujemaa [45], reduction of image signature dimensions was achieved using Principal Component Analysis (PCA). However, the use of PCA in the reduction of the dimensionality of image representations may lead to significant loss of pattern information due to its linear approach which may be inadequate for images [27].Zhang et al. [32]also proposed an improved hierarchical clustering: Graph Degree Linkage (GDL), which replaces the high dimensioned representation of each image with a K-Nearest Neighbour (KNN) graph developed by analysing the analysing the indegree and outdegree affinity between clusters using thus achieving non-linear dimension reduction before the Agglomerative clustering [32].

In the works of Kim et al. [41] and Huang et al. [3] spectral clustering was adopted as a means of providing the non-linear dimension reduction needed in the categorisation of images. While Kim et al. achieved non-linear dimension reduction via simple graph partition and link analysis, Huang et al. [3] introduced the application of hyper-graph partitioning to representing both local and global similarities among unlabelled images before the application of spectral clustering. Although spectral clustering method is capable of categorising any given image set irrespective of the complexity of the image signature, it favour compact and coherent groups over heterogeneous data collections and produce highly intricate clusters, which do not delineate separation between clusters [2]. It also require the calculation of an n2 order Pair-wise distances (where n is the size of the dataset) making the computation required for the procedure very high [2], which can make its application in the classification of a large image dataset set difficult. Another problem is its

reliance on the K-means algorithm, which also means that there is the need for prior knowledge of the number of categories present in the dataset.

Alternatively, Topic-based model can be used for the same purpose. Recently, the application of Topic-based model in semantic labelling has been generating some research interest due to its ability to capture image semantic contents while achieving dimension reduction [38, 46]. Although Topic-based model clustering is rated above centre-based techniques such as K-means clustering for unsupervised image categorisation [3], its classification accuracy is affected by the use of order-less BOV image representation [19, 21, 38, 22]. Topic-model based clustering can be improved through the inclusion of spatial information of visual words during BOV modelling [3, 21, 38]. Verbeek et al. [21] improved PLSA classification accuracy from 78.5% to 80.2% using local Markov Random Field (MRF). The same approach was used by Xu et al. in the unsupervised image classification using LDA. The spatial pyramid by Lazebnik [22] is another alternative for the reduction of spatial incoherency during the use of Topic modelling for image categorisation.

In general, all Mixture modelling clustering approaches are memory-based methods, the model is often built using the entire dataset, and the fitting is done at evaluation or prediction time, which makes this unsupervised approach naturally unsuitable for many real-time applications [33]. The PLSA/KNN approach of Bosch et al. [12] in which the authors built a PLSA Simplex using a fraction of the image collection as training images, after which any image to be classified is fitted to the simplex using Kullback-Leibler divergence is a likely solution to this drawback of mixture models. However, a disadvantage of this approach is that its use of KNN approach introduces the need for labelled training samples.

In general, the applicability of topic modelling to unsupervised classification can be further enhanced through the development of a method for determining the number of latent topics required for an efficient implementation, and the establishment of the relationship between latent topics and semantic contents (objects and locations).

4.3. Experimentation Protocol / Results

The accuracy of a typical unsupervised classification is determined by counting the number of image classification that matches the ground truth [3]. The performance of an unsupervised image categorisation process can also be displayed in details using a confusion table, where the overall performance is determined by the average value of the diagonal entries of the confusion table [3, 12, 41].In general, the main goal of an unsupervised image categorisation process is the allocation of each image of a dataset to one of a number of categories. However, Datta et al. [2] identified the unknown number of categories and the unknown nature of the categories in an image collection to be the two main challenges to the implementation of unsupervised image classification [2]. This sub-section examines how various implementations of unsupervised classifications have responded to these challenges.

Kim et al. [41] adopted the experimental protocol developed by Graum and Darrell [42] in evaluating their proposed unsupervised image categorisation. The experiment involves running ten iteration of the proposed algorithm on six object classes from Caltech101 (Airplanes, Cars, Faces, Motorbikes, Watches and Ketches) and three object classes from TUD/ETHZ (Giraffes, Motorbikes, Cars).The samples of images in the 6-chosen Caltech-101 classes are shown in Figure 3. For each iteration, the algorithm randomly picked 100 and 75 images from each object category in the Caltech101 and TUD/ETHZ datasets respectively. The experiment recorded an average of 98.55%, 97.30% and 95.42% for 4, 5 and 6 object Caltech101classes respectively, and recorded 95.47% over the TUD/ETHZ dataset. In a comparative experiment, Huang et al. [3] recorded 98.55%, 97.38%, and 96.05% under similar condition, which confirms the superiority of the ROI/hyper-graph partitioning technique over the simple graph partitioning.

Huang et al. [3] extended their experiment to 4, 8, and 12 randomly selected object classes from the entire caltech101 and caltech256 dataset. 100 iterations over Caltech101 revealed 95.8%,

86.2% and 71.5%, while in the case of caltech256 87.7%, 77.1% and 64.3% were recorded. While this result confirms the superiority of the proposed hypergraph based algorithm over chosen baseline unsupervised image classification based on affinity propagation, normalised cut, and K-centre, it also demonstrates a reduction in classification accuracy as the size of the dataset increases, and the increase in complexity of the images in the collection. Although Huang et al. [3]experimented further using PASCAL VOC2008 which recorded 81.3%, 77.2% and 69.3%, a result similar to what was recorded with caltech256 [3], this study recognises the increasing popularity of Caltech-101 and Caltech-256 datasets, and considers them to be a means of providing adequate challenge to all object recognition based experiments.

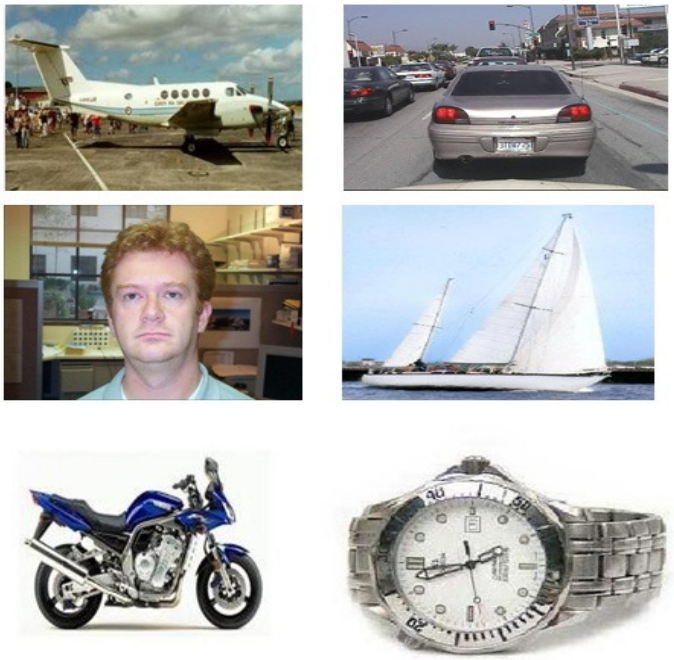

FIGURE 3: Sample images from the 6-categories chosen from Caltech-101 by Kim et al. [41] for the evaluation of the proposed unsupervised classification framework.

Zhang et al. [32] experimented with the proposed algorithm on object image databases (COIL-20 and COIL-100), hand-written digit databases (MNIST and USPS), and facial image databases (Extended Yale-B, FRGC ver2.0), and compared the GDL's performance to k-medoids (k-med), average linkage (Link), graph-based average linkage (GLink), normalized cuts (NCuts), NJW spectral clustering (NJW-SC), directed graph spectral clustering (DGSC), self-tuning spectral clustering (STSC) and Zell. The proposed algorithm demonstrated better robustness to noise and higher speed than the other algorithms. It also successfully implements unsupervised classification without the need for prior knowledge of the number of inherent categories while basing its categorisation on object matching thus accounting for the nature of each cluster [32].
This paper uses Figure 4 to illustrate the result of an evaluation of GDL carried out as part of its study. Although the average accuracy is approximately 85%, the graph indicates that the accuracy of the algorithm's classification reduces as the size of the experimental dataset increases, which may discourage its application to large image datasets. This paper considers further experiments on the object matching capability of GDL (and AGDL) using Caltech-101 and Caltech-256 to be necessary so as to establish a common evaluation environment with other recent works.

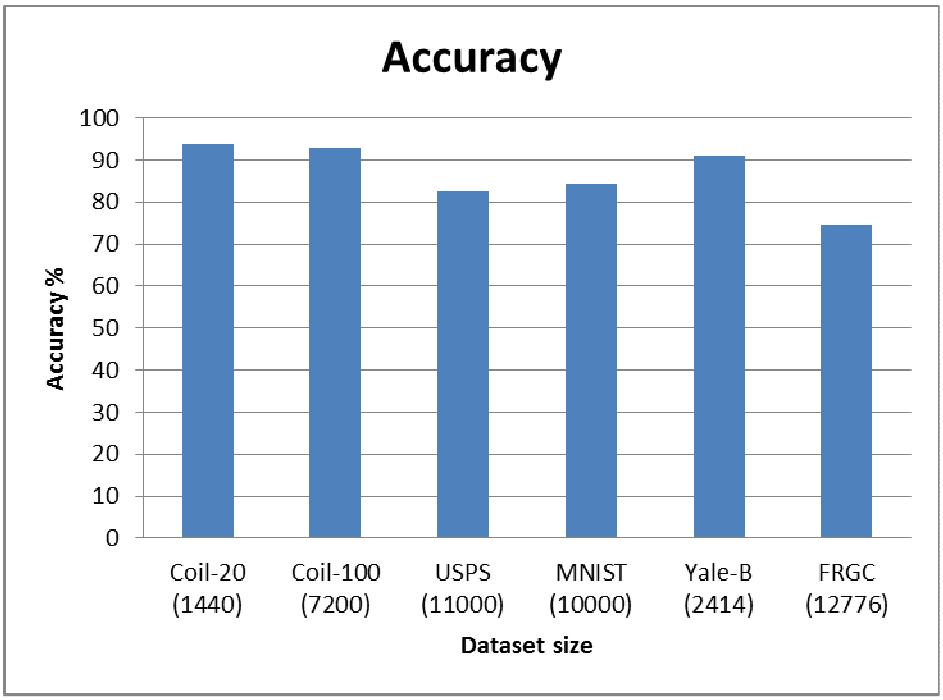

FIGURE 4: A summary of GDL classification accuracy showing variation in accuracy with the size of dataset.

Bosch et al. [12] evaluated their classification algorithm on three different datasets used in the supervised classification of Vogel and Schiele [47] and Oliva and Torralba [48], and the semisupervised classification of FeiFei and Perona [49]. The image collection used by Oliva and Torralba (OT) [48] has 8 categories, Vogel and Schiele (VS) [47] has 6 categories, while FeiFei and Perona (FP) [49] has 13 categories. Using 100 randomly selected images from the OT dataset, the authors determined that the optimum parameters for the experimentation are V = 1500, Z = 25, K=10 and M = 10, (where V= number of visual words Z= the number of topics, K= the number of neighbours for KNN and M = the number of pixels between cell locations during Dense-SIFT extraction), and the mean accuracy and standard deviation on the OT dataset (images of Natural and Man-made scenes) are of 84.78% and 1.93% respectively when dense colour SIFT was used as the feature descriptor.

The authors used the same parameters values on the OT, VS and FP datasets. For the OT dataset, they used approximately 200 images per category for both training and testing. In the case of the VS, they used 350 images per category for both training and testing, while a total of 1344 images were used for the FP dataset for training [12]. Despite the fact that Bosch et al. [12]training is unsupervised, the proposed algorithm outperforms all of the previous methods with 85.7% (against the previously recorded 74.1%) and 73.4% (against the previously recorded 65.2%) accuracies over VS and FP respectively, with the best classified scenes being highway and forest with 95.61% and 94.86% respectively.

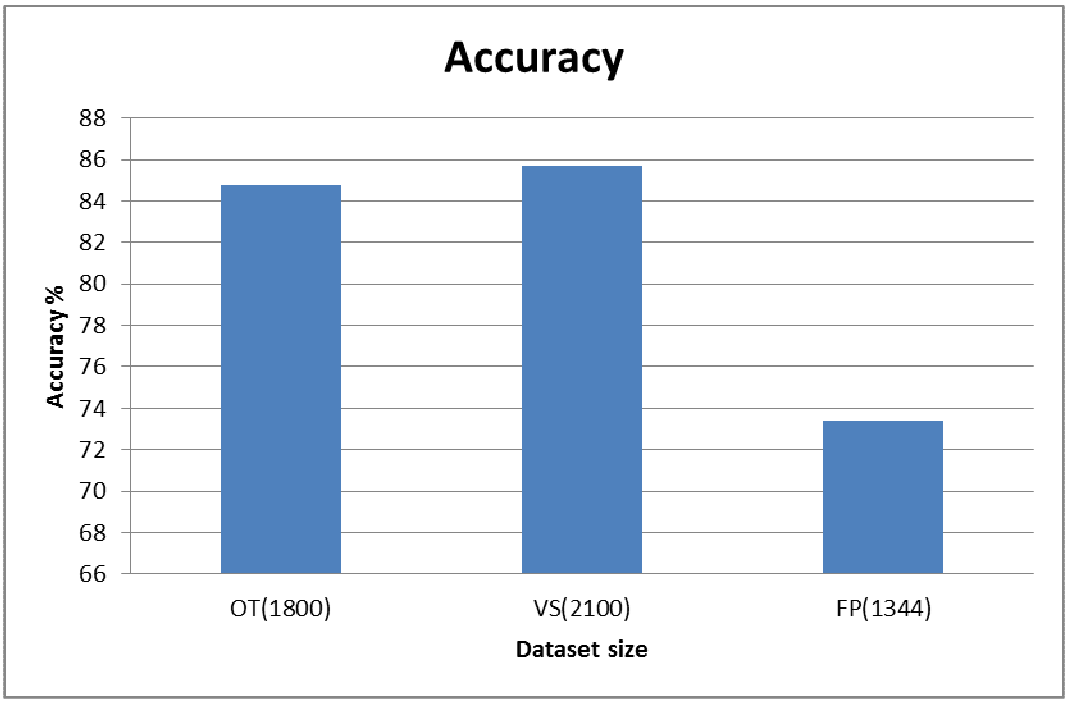

FIGURE 5: A summary of PLSA/KNN classification accuracy showing variation in accuracy with the size of dataset.

Bosch et al. [12] attributes the improved performance to the use of better features during scene categorisation, especially those features representing objects. This study recognises the ability of the combination of PLSA/KNN to implement unsupervised classification without prior knowledge of the number of inherent semantic group. This study also uses Figure 5 to demonstrate that the accuracies recorded through the use of PLSA/KNN combination in the unsupervised classification of images may not be responsive to the size of the dataset, making it a more suitable option than GDL for the categorisation of a large image collection (1000 and above). However, Figure 4 and Figure 5 do not offer a conclusive proof of the PLSA/KNN's superiority over GDL, therefore, there is a need to compare the two algorithms using Caltech-101 and Caltech-256 datasets, with emphasis on object recognition based unsupervised classification. This study also suggests an investigation into the effect of spatial incoherency (due to BOV modelling) on the PLSA/KNN combination.

5. FUTURE WORKS

The unsupervised classification of images offers a variety of opportunities as a solution to some problems in artificial intelligence. One of these problems is the elimination of the semantic gap present in CBIR via automatic annotation of images of a collection [2, 4, 50]. Wang et al. [4] explained that image retrieval researches are currently moving towards semantic-based image retrieval due to this presence of semantic gap in CBIR which has rendered its performance unsatisfactory.

Jeon et al. [51] proposed an automatic approach for the annotation of images with the aim of achieving convenient content based image retrieval. This approach depends on a supervised learning process which involves identifying common blobs from a set of labelled training images. However, like all categorisation based on supervised learning, obtaining adequate quality and quantity of labelled training images is a major challenge for this approach [51]. Therefore there is the need to look in the direction of unsupervised image categorisation.

The categorisation technique proposed by Bosch et al. [24] present a viable option for automated image annotation with the aim of eliminating semantic gap from an image retrieval process due to its use of PLSA which attempts to identify latent topics. However, the use of this technique requires labelled training samples due to the inclusion of KNN in the model, therefore there is a need for a research into a completely unsupervised but related technique. There is also a need for a research that clearly establishes the relationship between PLSA latent topics and semantic objects present in the image collection, these researches will enhance the use of unsupervised categorisation technique based on PLSA in the elimination of semantic gap from image retrieval processes.

6. CONCLUSION

Until recently, most research attention in image retrieval has been focused on feature extraction and similarity computation [2]. More recently, the need to minimise or totally eradicate the semantic gap from image retrieval systems has directed research efforts towards Semantic Image Retrieval in which the semantic gap is minimised through semantic labelling [4].Due to its ability to categorise images without the need for training samples, unsupervised image categorisation has the potential to facilitate convenient annotation of images in a large collection. Although, non-parametric clustering techniques are simple and intuitive, their direct application to a large image database is limited because they are not very suitable for clustering high-dimensional data [1, 33].

The use of image descriptive statistics via parametric clustering enables the capturing of important information about a given dataset [1]. This is especially so for the Topic-based model such as PLSA, that captures the relationship between visual-words and the frequency of their appearance on images. Hence it can be instrumental in matching low-level features to high-level semantics; thereby supporting Semantic labelling of images. This paper also recognises the ability of PLSA/KNN combination proposed by Bosch et al. [12] and the hierarchical clustering-based GDL proposed by Zhang et al. [33] to implement unsupervised image categorisation based on the nature of inherent groups within the image collection without the need for prior knowledge of the number of categories within the collection, therefore recommends a detailed investigation and comparison of their object recognition-based categorisation abilities using the increasingly popularCaltech-101 and Caltech-256 datasets. Such research may provide a more suitable means of mapping low-level features to high level semantics than existing methods for the elimination of the semantic gap in image retrieval processes.

7. REFERENCES

[1] 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.

[2] 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,, p. Article 5, Apri 2008.

[3] 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.

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

[5] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, January 2004.

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

[7] M. Guerrero, "A Comparative Study of Three Image Matcing Algorithms: Sift, Surf, and Fast," Utah State University, Utah, 2011.

[8] N. Khan, B. McCane and G. Wyvill, "SIFT and SURF Performance Evaluation against Various Image Deformations on Benchmark Dataset," in International Conference on Digital Image Computing: Techniques and Applications, Noosa, 2011.

[9] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," in Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference, Washington, 2004.

[10] L. Juan and O. Gwun, "A Comparison of SIFT, PCA-SIFT and SURF," International Journal of Image Processing, vol. 3, no. 4, pp. 143-152, 2008.

[11] C.-X. Liu, J. Yang and H. Huang, "P-SURF: A Robust Local Image Descriptor," Journal of Information Science and Engineering, vol. 27, pp. 2001-2015, January 2011.

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

[13] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," INRIA, Montbonnot, 2004.

[14] J. Brookshire, "Person Following using Histograms of Oriented Gradients," iRobot Corporation, Bedford, 2009.

[15] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in Computer Vision (ICCV), 2011 IEEE International Conference, Barcelona, 2011.

[16] A. Alahi, R. Ortiz and P. Vandergheynst, "Fast Retina Keypoint (FREAK)," in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference , Providence, RI, 2012.

[17] C.-F. Tsai, "Bag-Of-Words Representation in Image Annotation: A Review," International Scholarly Research Network, vol. 2012, pp. 1-19, 2012.

[18] A. G. Faheema and S. Rakshit, "Feature Selection using Bag-Of-Visual-Words Representation," in Advance Computing Conference (IACC), 2010 IEEE 2nd International , Patiala, 2010.

[19] P. Tirilly, V. Claveau and P. Gros, "Language Modelling for Bag-of-Visual Words Image Categorization," IRISA, Rennes, 2008.

[20] Y. Zhang, Z. Jia and T. Chen, "Image Retrieval with Geometry-Preserving Visual Phrases," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference, Providence, RI, 2010.

[21] J. Verbeek and B. Triggs, "Region Classification with Markov Field Aspect Models," in Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference, Minneapolis, MN, 2007.

[22] 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.

[23] 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.

[24] A. Bosch, A. Zisserman and X. Munoz, "Representing shape with a spatial pyramid kernel," in CIVR, Amsterdam, 2007.

[25] Y. Bai, L. Guo, L. Jin and Q. Huang, "A novel feature extraction method using Pyramid Histogram of Orientation Gradients for smile recognition," in Image Processing (ICIP), 2009 16th IEEE International Conference, Cairo, 2009.

[26] Z. Zhong and G. Shen, "Facial Emotion Recognition Using PHOG and a Hierarchical Expression Model," in Intelligent Networking and Collaborative Systems (INCoS), 2013 5th International Conference, Xi'an, 2013.

[27] L. Zisheng, J. Imai and M. Kaneko, "Facial-component-based bag of words and PHOG descriptor for facial expression recognition," in Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on, San Antonio, 2009.

[28] Q. Wang, "Kernel PCA and its Applications in Face Recognition and Active Shape Models," Rensselaer Polytechnic Institute, New York, 2011.

[29] B. Sch olkopf, A. Smola and K.-R. Muller, "Kernel Principal Component Analysis," Max-Planck-Institute, Tubingen, 1999.

[30] M. El Agha and W. M. Ashour, "Efficient and Fast Initialization Algorithm for Kmeans Clustering," I.J. Intelligent Systems and Applications, vol. 1, pp. 21-31, 2012.

[31] M. Seetha, I. V. Muralikrishna, B. L. Deekshatulu, B. L. Malleswari, Nagaratna and P. Hedge, "Artificial Neural Networks and Other Methods Of Image Classification," Journal of Theoretical and Applied Information Technology, pp. 1039-1053, 2008.

[32] M. Beale and D. Howard, Neural Network Toolbox, Natick: The Mathworks, 2002.

[33] W. Zhang, X. Wang, D. Zhao and X. Tang, "Graph Degree Linkage: Agglomerative Clustering on a Directed Graph," Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, 2012.

[34] P. K. Mallapragada, R. Jin and A. Jain, "Non-parametric Mixture Models for Clustering," Michigan State University, East Lansing, 2010.

[35] T. Hoffman, " “Probabilistic Latent Semantic Analysis”," in Uncertainty in Artificial Intelligence, Stockholm , 1999.

[36] J. Liu, D. Cai and X. He, "Gaussian Mixture Model with Local Consistency," in Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10), 2010.

[37] C. Tomasi, "Estimating Gaussian Mixture Densities with EM," 2004.

[38] D. M. Blei, Y. N. Andrew and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[39] S. S. S. Mole and L. Ganesan, "Unsupervised Hybrid Classification for Texture Analysis Using Fixed and Optimal Window Size," International Journal on Computer Science and Engineering, vol. 2, no. 9, pp. 2910-2915, 2010.

[40] T. T. Duong, J. H. Lim, H. Q. Vu and J. P. Chevallet, "Unsupervised Learning for Image Classification based on Distribution of Hierarchical Feature Tree," in Research, Innovation and Vision for the Future, 2008. RIVF 2008. IEEE International Conference, Ho Chi Minh, 2008.

[41] G. Kim, C. Faloutsos and M. Hebert, "Unsupervised Modeling of Object Categories Using Link Analysis Techniques," Carnegie Mellon University, Pittsburgh, 2007.

[42] K. Grauman and T. Darrell, "The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features," in IEEE International Conference on Computer Vision, Beijing, 2005.

[43] S. Todorovic and N. Ahuja, "Extracting Subimages of an Unknown Category from a Set of Images," in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference , New York, 2006.

[44] T. W. Lee and M. S. Lewicki, "Unsupervised Image Classification, Segmentation, and Enhancement Using ICA Mixture Models," IEEE TRANSACTIONS ON IMAGE PROCESSING, vol. 11, no. 3, pp. 270-279, 2002.

[45] B. Le Saux and N. Boujernaa, "Unsupervised Robust Clustering for Image Database Categorization," in IEEE Pattern Recognition 2002 Proceedings, 2002.

[46] G. Passino, I. Patras and E. Izquierdo, "Aspect coherence for graph-based semantic image labelling," IET Computer Vision, vol. IV, no. 3, p. 183–194, 2010.

[47] J. Vogel and B. Schiele, "Semantic Modeling of Natural Scenes for Content-Based Image Retrieval," International Journal of Computer Vision, vol. 72, no. 2, pp. 133-157, 2007.

[48] A. Oliva and T. Antonio, "Modelling the shape of the scene: A holistic representation of the spatial envelope," International Journal of Computer Vision, vol. 3, no. 42, pp. 145-175, 2001.

[49] L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference , San Diego, 2005.

[50] D. Zhang, M. Islam and G. Lu, "A review on automatic image annotation techniques," Patterrn Recognition, vol. 45, no. 1, pp. 346-362, 2012.

[51] J. Jeon, V. Lavrenko and R. Manmatha, "Automatic Image Annotation and Retrieval using CrossMedia," in ACM Special Interest Group on Information Retrieval (SIGIR), Toronto, 2003.