

# Visual and semantic context modeling for scene-centric image annotation

Mohsen Zand<sup>1</sup> · Shyamala Doraisamy<sup>1</sup> ·  
Alfian Abdul Halin<sup>1</sup> · Mas Rina Mustaffa<sup>1</sup>

Received: 14 August 2015 / Revised: 10 February 2016 / Accepted: 28 March 2016 /  
Published online: 6 April 2016  
© Springer Science+Business Media New York 2016

**Abstract** Automatic image annotation enables efficient indexing and retrieval of the images in the large-scale image collections, where manual image labeling is an expensive and labor intensive task. This paper proposes a novel approach to automatically annotate images by coherent semantic concepts learned from image contents. It exploits sub-visual distributions from each visually complex semantic class, disambiguates visual descriptors in a visual context space, and assigns image annotations by modeling image semantic context. The sub-visual distributions are discovered through a clustering algorithm, and probabilistically associated with semantic classes using mixture models. The clustering algorithm can handle the inner-category visual diversity of the semantic concepts with the curse of dimensionality of the image descriptors. Hence, mixture models that formulate the sub-visual distributions assign relevant semantic classes to local descriptors. To capture non-ambiguous and visual-consistent local descriptors, the visual context is learned by a probabilistic Latent Semantic Analysis (pLSA) model that ties up images and their visual contents. In order to maximize the annotation consistency for each image, another context model characterizes the contextual relationships between semantic concepts using a concept graph. Therefore, image labels are finally specialized for each image in a scene-centric view, where images are considered as unified entities. In this way, highly consistent annotations are probabilistically assigned to images, which are closely correlated with the visual contents and true semantics of the images. Experimental validation on several datasets shows that this method outperforms state-of-the-art annotation algorithms, while effectively captures consistent labels for each image.

**Keywords** Automatic image annotation · Visual diversity · Mixture model · Visual context · Semantic context

---

✉ Mohsen Zand  
m.mohsen.zand@ieee.org

<sup>1</sup> Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

## 1 Introduction

As multimedia content is explosively expanding where more visual information is available in digital archives, the need for effective image retrieval has greatly increased. Image retrieval mainly relies on captions and descriptions that describe a particular image [49]. In an ideal world where all images are properly annotated, image retrieval should be a near-solved problem through application of mature text retrieval techniques. However, since annotation involves cumbersome and rigorous manual labor, this utopia might be virtually impossible. It also causes errors, inconsistencies and subjectivity (especially for large collections). This has led researchers to explore automated approaches to solve (or partially alleviate) the image annotation process. Automatic image annotation (AIA) hence attempts to automatically identify and/or discover keywords to describe the contents of images [57]. Although many existing AIA algorithms can be found in the literature [37, 39, 57, 62], most of their performances are not entirely satisfactory. This has motivated researchers to further explore possible AIA solutions.

AIA algorithms can be categorized into two major groups. The first group includes AIA approaches that assign labels based on global image features [10, 52]. Such approaches allow straightforward indexing and retrieval since the annotations cover general terms. However, the main limitation is their failure to consider the fact that a single image can belong to multiple categories. This is hence less satisfactory for users with more complex queries that demand more detailed semantic content.

The second AIA category on the other hand, includes approaches that utilize local features to annotate visual objects with keywords. These approaches explore correlations among labels and local patches or segmented regions. Therefore, image annotation is treated and modeled as an image classification task, where each label is predicted using a multiclass classifier. Typical methods [6, 9, 43] usually learn a generative/discriminative classifier from training data to discover a mapping function from extractable low-level features (from the images or regions) to semantic concepts.

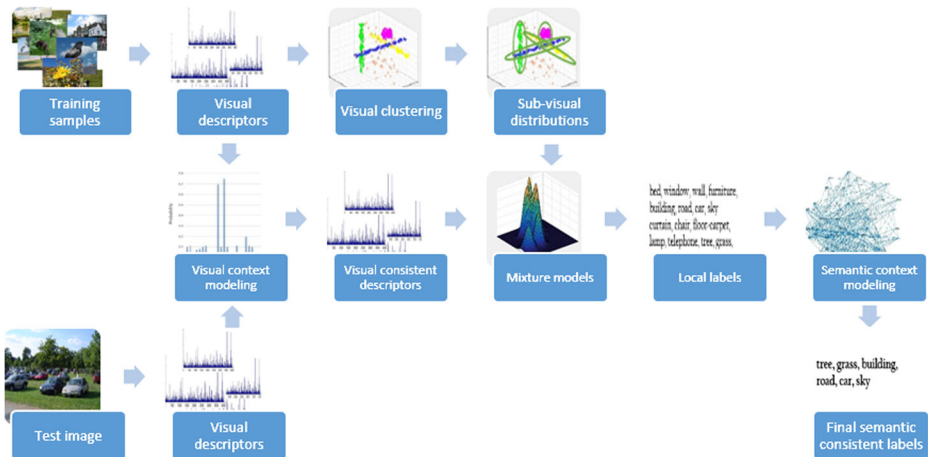
Practically, the second approach would be a more preferred strategy to conduct AIA. Although several multi-label AIA techniques have been recently proposed in the literature, it seems that the relationships among visual appearances and the semantic concepts are multi-aspect and intricate in large-scale image environments [31]. Specifically, visual diversity (concept polymorphism) and semantic confusion (visual polysemia) are two main issues being reported [48, 63, 65]. Visual diversity is due to the reason that a certain concept can have different visual appearances under different circumstances (e.g. orientation, scale, lighting condition, etc.). Consequently, an object representing one single concept can possibly be characterized by different visual features in different images. This is also coupled with the facts that current segmentation methods are not mature enough to robustly partition objects from images, and with the projection of 3D real scenes to 2D space, only one view of the object appearance is captured. On the other hand, semantic confusion exploits the apparent fact that a visual pattern usually shares different semantic meaning as it can occur in many concepts. The incomplete and noisy feature extraction process, and inconsistent visual features also lead to extremely ambiguous visual patterns. This means that it is difficult to assign an exact semantic label without contextual information. In addition to these problems, most current methods rely on automatic segmentation algorithms that are usually unable to decompose images into their respective semantic regions [5]. Hence, learning a single concept model solely based on the visual appearance is very challenging.

Typical methods attempt to discover correlations between visual descriptions and semantic concept [47, 53]. The representation commonly used in these approaches is the bag-of-visual words (BoW) model [13], which characterizes each image as a histogram of ‘visual words’ in a visual dictionary. The visual words are computed using an unsupervised vector quantization algorithm such as k-means on the low-level features of local patches or key points. Despite the scalability and efficiency of this model, one main drawback is feature mismatching which is due to the ambiguity of the visual words. To minimize the quantization loss between local features and visual words, supervised dictionary learning has also been proposed in the literature [18, 28]. In [18], Gao et al. proposed a weakly supervised dictionary learning method which incorporated cheaply available visual features, and iteratively refined the quantization process and feature-to-visual words mappings. Nevertheless, it is implicitly assumed that all the image patches related to a semantic concept are visually correlated. As a result, the generated low coherence visual words are not expressive of the semantic concepts. Moreover, the distance measures used for the clustering of the high dimensional visual feature space exhibit a more sensitive nature [45].

This paper proposes a novel approach, called visual- and semantic-consistent image annotation (VSCIA) to tackle the above mentioned issues. The visual diversity problem is alleviated by grouping all the samples in the same class into an unknown number of sub-visual distributions which are visually coherent. This is specifically achieved using a clustering algorithm which can efficiently partition the high-dimensional image data in each semantic class. Subsequently, the sub-visual distributions of each semantic class form a mixture model for that class. As a result, non-ambiguous descriptors can be associated with image labels through these mixture models. Inspired by the fact that image annotations are visually and semantically consistent for each image, the contextual knowledge is incorporated into the model in both visual and semantic levels. This can not only discern the actual visual contents of the images but can also reduce the semantic confusion and generate more coherent annotations. Chiefly, the visual context is learned using a probabilistic Latent Semantic Analysis (pLSA) model which characterizes the joint distributions between images and visual contents via latent visual topics. Therefore, all images share the same set of visual topics which provide context to ambiguous descriptors in each image and thus, maximize the likelihood about the actual visual content. This helps to assign relevant local labels to the visually consistent local descriptors of an untagged image. Yet, the scene-centric representation of the images enforces to investigate the global consistency through local labels of each image. This can be achieved by the proposed semantic context modeling, where a concept graph exploits the correlations between semantic concepts. It also enables to probabilistically rank the semantic concepts with regard to the image scene. Furthermore, highly consistent annotations can be associated to each image, which are closely correlated with its visual contents and true semantics.

The overall framework of the proposed model is shown in Fig. 1. In the training phase, the visual clustering algorithm generates an unknown number of sub-visual distributions for each semantic class. The visual-consistent descriptors of the samples in the sub-visual distributions, which are derived from the visual context model are then used to train the mixture models. Subsequently, the global semantic context is learned for the final labeling an untagged image in the testing phase.

The contributions of this work are twofold. First, a visual clustering algorithm is proposed for high dimensional image data, which can partition diverse visual samples of a semantic concept into unknown number of coherent sub-visual distributions. Second, the proposed



**Fig. 1** Schematic illustration of the proposed VSCIA system

contextual modeling can reduce the semantic confusion by leveraging context in both visual and semantic levels, and hence perceiving each image in its particular scene. Specifically, the visually consistent descriptors of each image scene determined by the visual topics in a pLSA model are associated to their labels through mixture models, and the semantically consistent labels explored in a concept graph are presented in the final annotations.

The reminder of this paper is organized as follows. Section 2 explains the related work. Clustering algorithm that deals with the diverse visual samples of the semantic concepts is presented in Section 3. In Section 4, we introduce the scene-centric image annotation by context modeling. Experimental results and discussions are given in Section 5. Finally, some concluding remarks of this paper and the future work are presented in Section 6.

## 2 Related work

As a popular research topic in recent years, many research efforts have been devoted to the problem of image annotation. It is usually cast into a machine learning strategy to combine multiple sources of information in order to associate visual features with semantic concepts. Despite interesting advances, it suffers from two main issues: visual diversity (concept polymorphism) and semantic confusion (visual polysemia).

To address the visual diversity problem, the major challenge is to find class models that are invariant enough to incorporate different visual variations and yet discriminative enough for broad semantic classes. Many works have been conducted in combining multiple visual words and modeling their visual relationships. Zheng et al. [65] presented a visual synset as a semantic-consistent cluster of visual words. They used distributional clustering based on information bottleneck principle to group the visual words in visual synsets. Although this technique can partially retrieve visually diverse samples of the same semantic class, the visual synsets do not take the spatial contexts among visual words into consideration. Moreover, influence of the number of the classes on the semantic inferences of the visual synsets, and the final classification are not investigated. In [31], Li et al. incorporated the BoW model into a Vcept (visual appearance-to-semantic concept) image representation as a hierarchical

representation of image semantics. In their model, membership distribution between each visual word and semantic concepts was established. Therefore, the problem of visual diversity was addressed by learning the probability relationships between one certain concept and all the visual appearances. However, this method uses learning algorithm via mixed-norm regularization optimization which is an expensive task. It also depends on the purified image training database which is not easily available.

Several studies have proposed to capture the existing spatio-contextual relationships in natural images, and reduce the semantic confusion. Lazebnik et al. [29] calculated the distribution of the visual words at multi-spatial resolutions, which introduced some information about the distributions of the visual words surrounding the region of interest. They then utilized spatial pyramid matching to measure the similarity of the images. However, wrongly representing the local regions in a specific spatial resolution may cause errors in context representation. Tirlly et al. [51] used simple spatial relations between visual words in their visual sentences. They employed pLSA model to remove the noisiest visual words in a language modeling approach. Vogel and Schiele [54] constructed a semantic vocabulary by manually associating local patches to concepts for semantic modeling of natural scenes. They divided images into regions, predicted the categories of these regions, and used normalized histograms of the concepts as concept-occurrence vectors for a global image representation. Their work was based on the idea that the meaning of an image can be described using the meaning of its constituent visual words. Rasiwasia and Vasconcelos [42] presented a holistic model to compute the co-occurrence statistics on the whole image without considering the local spatial relationships among features. They represented images in two levels of concept representation and semantic space. The concepts were modeled by formulating the probability distributions of the semantic multinomial which were derived from the images of each concept. The concept models were then used to represent images as vectors of posterior concept probabilities under these contextual concept models. In [17], word frequencies were integrated with the spatial neighborhood representation by considering each visual word as an item, and counting the number of items in a novel set of local patterns. However, the relevant and non-redundant constraints cannot be directly applied on the local patterns mining process.

More complex graph relations [3, 25, 64] were presented in the literature to model the visual and semantic representation. For instance, Zhao et al. [64] proposed a multi-graph learning model to handle the action retrieval problem. In each graph, they used the similarities between images in terms of visual features in different levels, and combined the multiple graphs in a regularization framework to exploit the complementation of various features by learning the optimized weights of each graph. Assari et al. [3] proposed a contextual approach to video classification based on generalized maximum clique graphs. They used co-occurrence of concepts and classified a video based on matching its semantic co-occurrence pattern to each class representation. Izadinia et al. [25] modeled the joint relationships between low-level events in a graph for the problem of complex event recognition. They used a graph structure to learn co-occurrence patterns of the low-level events in a latent SVM training procedure. Li et al. [32] proposed to model the relevance between an image and a tag as a pair-wise similarity in a unified space in their projective matrix factorization with unified embedding (PJMF) method. They constructed a correlation matrix from two low-dimensional latent representations for images and tags in the unified space, and used it to the social image retagging task. In [55], Wang et al. presented a complex graph clustering to address a high-level semantic image annotation based on hot Internet topics. They exploited the hot Internet topics through the modeling and clustering of the complex graph, and annotated the images by top keywords in

the corresponding hot topics. Nevertheless, these approaches rely on local relations between adjacent nodes, and thus ignores the long-term relationships. Markov Random Fields (MRF) and Conditional Random Fields (CRF) are also two common approaches to capture the spatial relationships among neighboring regions. Multi-scale CRFs were utilized in [23] in order to annotate regions using both local and global image features. Jiang et al. [26] proposed a Boosted CRF Concept Fusion (BCRF-CF) framework to model the inter-conceptual relationships and refine the detection results for each of the target concepts. It iteratively discovered related concepts through independent concept detectors. In [41], local detectors were used to assign primary object labels to segmented regions, and the labels were adjusted using a CRF. The extracted global features from the inter-class spatial relationships and locally related location features were utilized in [20] to annotate regions. It used CRF to include pairwise affinity preferences between neighboring pixels/regions. Ladicky et al. [27] proposed a hierarchical CRF model which integrated visual features and contextual priors over multiple image segmentations. Llorente et al. [36] used MRF to incorporate the statistical co-occurrences of quantized visual features and spatial relationships. However, these graph-based approaches have limitations in the sense that they are computationally expensive in training the high order graph structures. The complexity is also increased with the large number of classes. Therefore, the existing methods only consider very limited number of adjacent nodes in the graph structure.

The evidence in the literature suggests that contextual relations can provide higher order statistical information and enhance the discriminative power of features. These contextual features involve larger supporting regions than a single feature and hence can represent more complex structures in images. However, most of the existing methods compute the spatial co-occurrences from either the locally adjacent regions or the whole image. More importantly, these approaches only verify spatial consistency of features within local areas at one of the visual or semantic levels instead of the entire image scene. Although computationally efficient, they cannot capture the contextual relationships between all descriptors, and thus, they obtain limited performance improvement.

In order to overcome the visual diversity problem, a visual concept learning is proposed in this paper. It aims at the visual diversity problem using a clustering algorithm, where incoherent data in a visual cluster are forced to construct new clusters with the same semantic. Therefore, each visual cluster can be shaped by visually consistent samples for efficiently modeling a sub-visual distribution. This can result in the minimization of the intra-class diversity, while the visual discrimination between clusters can be maximized. In addition, it can also handle the curse of dimensionality in image descriptors. Using mixture models, these visual clusters can be associated with semantic classes in order to locally annotate images. Although the local label assignment can be enhanced, the image annotations are not expressive enough to describe image scenes semantically. This is because the feature descriptors are extracted individually, and therefore, the visual and semantic consistencies of the global scene contexts are not taken into account.

In this work, a principled solution is proposed that can leverage visual and semantic contextual modeling to improve the limitations of the traditional methods, and reduce the semantic confusion. It is inspired by the human perception in explaining the image content. The humans usually perceive the visual scene as a whole entity and without ambiguity with respect to the visual categories that the image belongs to, such as indoor or outdoor scenes. Coherent semantic terms are then matched the visual content to describe the image. Therefore, both the visual and semantic consistencies are necessary to be achieved across the entire image



scene in order to develop an efficient image annotation system. Correspondingly, the contextual relationships are exploited in both visual and semantic levels. Using pLSA, the visually relevant images are linked through latent visual topics which provide context to ambiguous descriptors in each image. A concept graph is also constructed from semantic concepts in order to find frequent (or repetitive) co-occurring concepts to explore the spatial dependency in semantic image data. Furthermore, our proposed model is different from the existing approaches since the relationships between images and their visual contents, the correlations of the semantic concepts and local image descriptors, and dependency among semantic concepts are leveraged efficiently. This can potentially boost the image annotations since non-ambiguous descriptors better map the visual space into the semantic space, and the coherent semantic meaning as a whole entity can be better preserved.

### 3 Clustering for visual diversity reduction

Most existing methods for AIA generate a visual codebook by grouping the low-level features extracted from training images into a predefined number of clusters, treat the center of each cluster as a visual word, and then annotate an unseen image by finding the closest entry in the codebook with the extracted features of the image. However, these methods rely on the assumption that the high-dimensional image features in each cluster are uniformly distributed in the Euclidean space. This assumption however, might not hold true for all cases because objects within the same semantic category may not be perceptually similar. This is the intra-class diversity which is inevitable in some semantic categories. Therefore, current AIA methods which assign the keywords based on the visual features without considering different viewpoints, poses and lighting conditions tend to fail.

To relieve the intra-class diversity, the training samples must be grouped based on their visual features. The success of the grouping procedure is not guaranteed since image visual descriptors contain heterogeneous features. It is desirable to integrate heterogeneous yet complementary feature descriptors to exploit more characteristics of images for discrimination. Multi-view learning or learning with multiple distinct feature sets can be used to handle this problem by considering the diversities of different views. In [34] for instance, multi-view Hessian regularization (mHR) was proposed to combine multiple Hessian regularizations obtained from multi-view features. Although it can explore the complementary properties of different features from different views, it is still challenging because of the difficulties in correlation discovery between multiple views. In [59], Xu et al. extended the theory of the information bottleneck to model the multi-view learning problem. They introduced the margin maximization approach to improve the code distance of the encoded examples. The multi-view data were then mapped into a new subspace as a compact representation of the original space. However, these methods concentrate on supervised or semi-supervised learning where a validation set is required. Specifically, in a semi-supervised learning method, some labeled examples are given to predict the labels of unseen examples. In this paper, we focus on multi-view clustering, which is much more difficult for lacking training data to guide the learning stage.

Note that in order to achieve uniform visual descriptors for a reliable labeling in our VSCIA method, the visual diversity must be tackled in each semantic class. Therefore, the training examples of each class with inherent visual variations are taken into account for diversity reduction. However, the training examples in the same category are labeled

identically which are in fact supposed to be unlabeled. In other words, there is no visually uniform labeled training examples in each semantic class to use supervised or semi-supervised learning approaches. In fact, diversity reduction is an unsupervised learning problem, where clustering methods are applied to organize data into groups of similar members in a collection of unlabeled data. However, the popular clustering methods such as K-means that are usually developed to handle uniformly distributed data (or single-view data) are often unreliable for high-dimensional multi-view image data. They also require the number of clusters to be known beforehand, which is usually hard to determine. In addition, it is noticeable that the multi-view structure of training examples in each semantic class forces the clusters to be irregular or intertwined. Therefore, the clusters exist either in a single view or combinations of views with different densities separated by less dense regions. This can be partially solved by density-based clustering methods such as DBSCAN [16] that rely on this kind of density estimation. However, these methods are based on a single global density threshold which cannot suitably characterize multi-view data with variation of densities. Likewise, although many hierarchical clustering approaches such as CURE [21] and OPTICS [2] do not constrain the shape of the clusters, and do not require predefined parameters, they are not satisfactory for diversity reduction in multi-view features of the semantic classes. These methods often use a distance matrix as their input and iteratively obtain a hierarchy of clusters, called dendrogram, to represent cluster relatedness. A partition can then be found by fixing a cut-off threshold on the dendrograms at a specific level. However, similar to the density-based methods, most hierarchical methods use a global density (cut) threshold through a hierarchical cluster representation. Moreover, multi-view features need a complex criterion for merging (or separating) the clusters specially when noise and outliers are present. To deal with the problems above and preserve the maximum information, an efficient clustering method is proposed, in which given a set of feature vectors labeled by an individual semantic concept, an optimum number of clusters is generated in an incremental manner. It can leverage the advantages of both hierarchical and density-based clustering algorithms. The proposed method focuses on the use of density-like information in the multi-view features, and obtains clusters of arbitrary shape while avoiding the problem of the global density threshold. It is hence a subspace clustering based on the HDBSCAN algorithm [8] that can efficiently handle the curse of dimensionality in image feature descriptors by using dimension reduction. Particularly, it is built on the idea that all the clusters can no longer be found in the entire feature space especially for the high-dimensional image data. In addition, it is likely that clusters lie in the subspaces due to the variation of the visual features for a specific class label. For instance, a specific feature of a semantic concept, such as color or shape of a car, may take totally different values in the feature space. This is the matter of subspace clustering algorithms which aim at finding all clusters in all subspaces using heuristic search techniques [7, 40]. However, these algorithms usually involve a lot of redundancy as they allow multiple cluster memberships. Instead, we propose a simple yet effective clustering in a top-down manner using the ‘noise’ concept. Starting with the entire feature space, the possible clusters are created. The remaining samples are considered as ‘noise’ which can be grouped in the further clusters in the reduced dimension space. This algorithm can find all the disjoint clusters hierarchically. The clustering hierarchy is achieved as a dendrogram from a minimum spanning tree (MST) which is capable of detecting clusters with irregular boundaries. Note that the only priori



knowledge required for finding the optimal clustering is the minimum number of samples to form a cluster, which is a classic smoothing factor in density estimates.

Let  $X = \{x_1, x_2, \dots, x_N\}$  denote the data samples, where  $x_i = \{\eta_1, \eta_2, \dots, \eta_d\}$  is a set of  $d$  different image features. The  $m_{pts}$  denotes the minimum number of samples in each cluster. The  $dis(x_p, x_q)$  stands for distance metric between  $x_p$  and  $x_q$ . The  $dis_{core}(x_p)$  is the distance from  $x_p$  to its  $m_{pts}$ -nearest neighbor. The mutual reachability distance is computed by  $dis_m(x_p, x_q) = \max\{dis_{core}(x_p), dis_{core}(x_q), dis(x_p, x_q)\}$ . The  $G_{mpts}$  is a complete graph where vertices are  $x_i, i = 1, \dots, n$ , and each edge is the mutual reachability distance between two respective connected vertices. The main steps of the proposed algorithm are summarized in Algorithm 1.

Algorithm 1: Clustering of high-dimensional image data by dimension reduction

---

Algorithm 1: Clustering of high-dimensional image data by dimension reduction

---

- 1-  $d_w = d, X_w = X$
  - 2- Iterate the following steps until  $d_w = 0$ .
  - 3- For  $\binom{d}{d_w}$  different subspaces, iterate the following steps.
  - 4- Compute the  $dis_{core}(x_i)$  for all  $x_i$  in the  $X_w$ .
  - 5- Compute the  $G_{mpts}$ .
  - 6- Compute MST of the  $G_{mpts}$ .
  - 7- Add self-edges with  $dis_{core}(x_i)$  as the edge weight to the vertices of the MST.
  - 8- Iteratively remove all edges from MST in descending order of edge weights.
  - 9- Assign appropriate labels to the connected components.
  - 10- Compute  $X_{noise}$  which are vertices without any edge.
  - 11- Set  $d_w = d_w - 1$ , and  $X_w = X_{noise}$ .
- 

Obviously, all the subspaces are considered by  $\binom{d}{d_w}$  which shows all the combinations of different visual features. Starting from the entire feature space by assigning  $d_w = d$ , the most similar samples are grouped into one cluster and removed from the clustering process in the following iteration. With the MST of the mutual reachability graph, the connected components and noise samples are identified and labeled at each level. The partitioning of the data is similar to DBSCAN [16] that runs Single-Linkage over the transformed space of mutual reachability distances, cuts the resulting dendrogram at a level of  $\varepsilon$ -radius, and treats all resulting singletons with the  $dis_{core}$  greater than  $\varepsilon$  as a “noise” class. In DBSCAN, the radius parameter should be predefined, which apparently, is difficult to estimate. Instead, different values of the radius are examined in this method by considering different density levels (threshold) in the cluster hierarchy. Note that clusters are defined as connected components in each density level, and boundary examples which their estimated density is below the threshold are noise. This top-down procedure efficiently produces a clustering tree that contains all disjoint clusters for each semantic category. This algorithm attempts to capture the coherence that exists in a subset of samples on a subset of image features. Therefore, the number of features used in the distance metric is different in each cluster. Using this algorithm, the most conceptually and visually similar samples can be collected in a cluster based on the multi-view features. The computational complexity of the proposed algorithm is  $O(dn^2)$  while the required memory is  $O(n^2)$  when the Prim’s algorithm is used to construct the MST (building the tree one edge at a time by adding the lowest weight edge), and the distance matrix is provided as input.

## 4 Scene-centric annotation

The proposed annotation technique is a multi-label multi-instance learning framework [60], which in addition to probabilistically assigning local labels, also models the contextual relationships at both visual and semantic levels.

As discussed in the previous section, visual-incoherent examples in a semantic class form new clusters of sub-visual distributions with the same semantic. Therefore, the correlation among a semantic class and its visual content exhibits a one-to-many relationship. To characterize this relationship, mixture models can be used, which probabilistically associate the sub-visual distributions with their corresponding semantic classes. Once the mixture models are built, they are utilized to reveal semantic labels of an unseen image based on its visual descriptors. However, the visual descriptors might be incomplete, noisy or ambiguous due to the lack of mature feature extraction, sensitivity to small errors in feature extraction, high dimensionality, and coarseness of the compositional descriptors. A promising solution is to consider an image as a unified entity, and integrate the contextual dependencies among visual elements of an image scene. This is inspired by the fact that a collocation of several visual concepts is likely to be much less ambiguous since the scene context is image-specific. Therefore, to be more robust to imperfection and unreliability of the visual descriptors, the global consistency is imposed on the visual descriptors of the same image. This can be achieved by topic modeling before the label assignment in the mixture models. In this work, a pLSA model is proposed to characterize latent visual topics between images and visual descriptors. These visual topics correspond to image patches with similar visual attributes, and they allow each image to be represented as a mixture of visual topics.

However, similar to the visual dependency in the visual context, the semantics conveyed by different labels for a specific image are actually correlated. Therefore, ignoring contextual relations among annotation words, and labeling the local image patches independently cannot achieve robust annotation results. Contextual knowledge is indeed embedded in the manual annotations, since humans usually annotate an image with a set of keywords with coherent semantic meaning as a whole entity (with respect to the visual content) rather than annotate each keyword one by one [57]. A straightforward strategy to obtain the contextual knowledge is through the co-occurrence frequency of pairs of concepts. It can be reasonably assumed that if two concepts are similar or related, their contextual environments will be equivalent, and they tend to appear in similar contexts. However, the contextual similarity is dependent on the concept distributions in the database [4]. To capture the contextual information, a concept graph is constructed between semantic concepts, where thicker edges represent more correlated labels. Using this structure, concepts with stronger relationships/higher correlations are taken into account, making the final labeling more consistent for each image.

### 4.1 Modeling visual context by pLSA

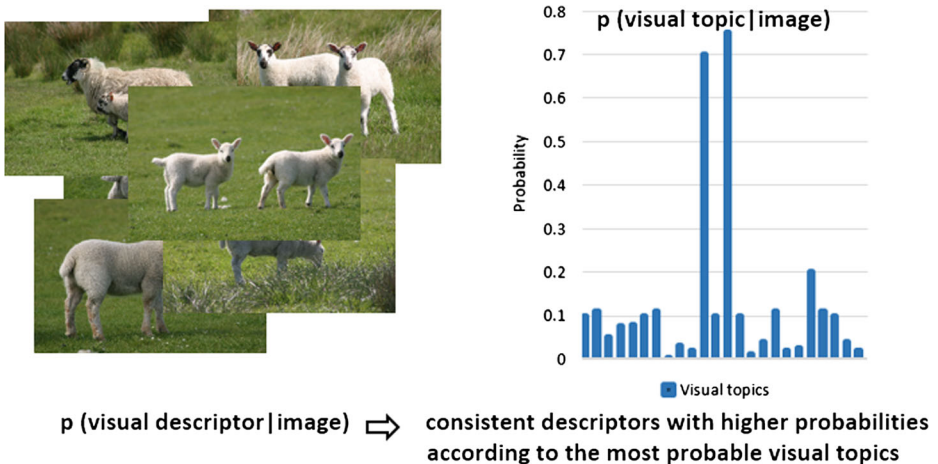
The proposed VSCIA model investigates the contextual correlations of each image at both the visual and semantic levels. As for the visual context of images, a pLSA model is used to analyze the consistency of the combination of the visual descriptors in each visual scene. Preserving the local consistence in the feature space has been recently approached in the literature [19, 34, 35]. In [35] for instance, Liu et al. proposed a multi-view Hessian discriminative sparse coding method to encode the intrinsic local geometry of data manifold for image annotation. However, as shown by Donoho and Grimes [14],

the Hessian-based methods exhibit undesired results in very high-dimensional data since they require estimations of second derivatives. In addition, although their method can leverage the local geometry of the data distribution, and utilize the complementary information of multi-view features, it is not a precise way for the local consistence of visual context, and visual disambiguation for semantic labeling. Instead, we show that co-occurrence embeddings relate statistical correlations of the local visual structures. Specifically, we aim at introducing a latent visual topic layer between two observable components (i.e. images and descriptors in our case). The goal is to find these latent visual topics corresponding to visually similar scenes that frequently occur in the dataset. The visual topics are shared across all the images, and hence they can exploit a rich context to ambiguous descriptors in each image.

The proposed model assumes that there are a limited number of visual descriptors denoted by  $V$ , and a database consisting of  $N$  images. The corpus of image documents is summarized in a  $V$  by  $N$  co-occurrence matrix, where  $n(v_i, I_j)$  denotes the number of occurrences of a visual descriptor  $v_i$  in image  $I_j$  for  $i = 1, \dots, V$  and  $j = 1, \dots, N$ . This model associates an unobserved visual topic variable  $z \in Z = \{z_1, \dots, z_k\}$  with each observation. A joint probability distribution  $p(v_i, I_j)$  of an image  $j$  is governed by the hidden conditional distribution of the visual context as:

$$p(v_i, I_j) = \sum_{k=1}^K p(z_k, I_j) p(v_i, z_k) \tag{1}$$

where  $p(z_k, I_j)$  is the probability of topic  $z_k$  occurring in image  $I_j$ , and  $p(v_i, z_k)$  is the probability of visual descriptor  $v_i$  occurring in a particular topic  $z_k$ . This model represents each image as a convex combination of latent visual topics [44]. In other words, a particular image is composed of a mixture of the visual topics corresponding to a specific visual scene. As shown in Fig. 2, the similar images share the same set of visual topics. In the visual-based latent space,  $p(\text{visual topic}|\text{image})$  denotes the degree to which an image can be represented using the corresponding latent visual topics. It is quite predictable that the mixture models applied on the consistent visual descriptors can generate more reliable local labels.



**Fig. 2** Similar images share the same set of latent visual topics, which their posterior probabilities aim at visual consistency through each image

## 4.2 Mixture models for labeling local image descriptors

The visual diversity problem can be alleviated by the clustering algorithm which partitions visually diverse samples from the same semantic class into coherent sub-visual distributions. Therefore, the diversity reduction leads to a one-to-many relationship between a semantic class and its visual content. This makes the label assignment more complex, where the common approaches that are based on the visual-to-semantic modeling [61] are unsuitable. Intuitively, this paper proposes to probabilistically associate the sub-visual distributions with their corresponding semantic classes using mixture models which can represent arbitrarily but finite number of densities. For an untagged image, visual-consistent image descriptors discovered in the pLSA model can be matched to the most relevant semantic classes in a mixture modeling scheme. The mixture models are learned for the semantic concepts in the training stage. Accordingly, the generated clusters of a semantic concept build a mixture model for that concept. A larger cluster with more samples possesses a higher likelihood of belonging to its relevant concept. In this context, cluster and component are used interchangeably because every mixture component is estimated using cluster samples. The centroid of the mixture components are referred to as codewords, which are generated by the quantization of the cluster samples [33].

The samples belonging to a given component are assumed to be drawn from a multivariate normal distribution  $N(\mu, \sigma^2)$  that is a better approximation to real data [30], where  $\mu$  is the map of the relevant codeword. The consistent visual descriptor  $v_i$  in image  $I_j$  is associated with the  $m$ -th concept containing  $\delta$  codewords based on the probability density function:

$$\phi(v_i|\lambda_m) = \sum_{n=1}^{\delta} \omega_n \left( 1/\sqrt{2\pi\sigma_n^2} \right) e^{\left( \frac{D(v_i, \mu_n)}{2\sigma_n^2} \right)} \quad (2)$$

where  $\lambda_m$  is the model of  $m$ -th concept,  $D$  is a distance metric, and  $\omega_n$  are the prior probabilities for clusters  $n=1, \dots, \delta$  (in  $m$ -th concept) with the constraint  $\sum_{n=1}^{\delta} \omega_n = 1$ . The priors are estimated by the percentage of the descriptors assigned to the codewords. The maximum likelihood (ML) method is used to estimate the parameters of each component [1].

Once the local labels are assigned, the label consistency in the image is determined using a concept graph. More precisely, the concept graph is constructed from the co-occurrences of the labels in all images, and then used to investigate the consistency and rank of the final label set for each image. The high ranked label is the most consistent label with the other labels.

## 4.3 Modeling semantic context by a concept graph

The contextual relationships between semantic concepts are modeled using a graph structure, which can provide a powerful tool to significantly improve the image annotation performance. This concept graph is a relational model consisting of concepts which are interlinked unidirectionally [22]. It is based on the idea that co-occurred concepts induce a similar scene. In other words, typical correlated concepts often co-occur in the same image. This resembles the manifold learning algorithms such as the Hessian regularized support vector machines [50] that assume relevant image pairs are derived from the uniform conditional distribution pairs. Nevertheless, in addition to the need of a large set of unlabeled samples, they are ideal for images containing a single object. Alternately, the proposed model constructs the concept graph using the co-occurrences of the concepts in the training set, and utilizes it to

probabilistically rank the most consistent labels to the given scene. It can also model the concept relationships more robust than the existing graph-based methods [20, 23, 26, 41] that only investigate very limited number of adjacent nodes. This is because our concept graph observes the image scene as a whole entity, and characterizes the relations between all the local labels.

Each node in the concept graph is a concept, and each edge reflects the co-occurrence relationship between the two connecting concepts. Let  $y_I = \{y_I^1, \dots, y_I^M\}$  denote the labels of image  $I$ , where  $M$  is the number of semantic concepts, and  $|y_I^m \cap y_I^p|$  denotes the co-occurrence frequency of the concept  $c_m$  with the concept  $c_p$  in image  $I$ . Correspondingly, the concept co-occurrences are used to decide whether an edge exists between two concepts. The edge weight between the two concepts  $c_m$  and  $c_p$  can thus be formulated as:

$$w_{m,p} = \begin{cases} f_{m,p}, & \text{if } f_{m,p} > \theta_m \text{ or } f_{m,p} > \theta_p \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where  $f_{m,p} = |y_I^m \cap y_I^p| + \dots + |y_N^m \cap y_N^p|$  is the overall co-occurrence frequency of two concepts  $c_m$  and  $c_p$  in all  $N$  images of the database, and  $\theta_m$  represents the average co-occurrence frequency for the concept  $c_m$  considering its neighboring concepts. In the experiments, an average co-occurrence frequency is set for each concept because specific co-occurrence frequencies represent varying importance for different pairs of concepts, and using several thresholds is helpful to extract more reasonable neighbors for different concepts. The final label set is generated using a sigmoid function defined as follows:

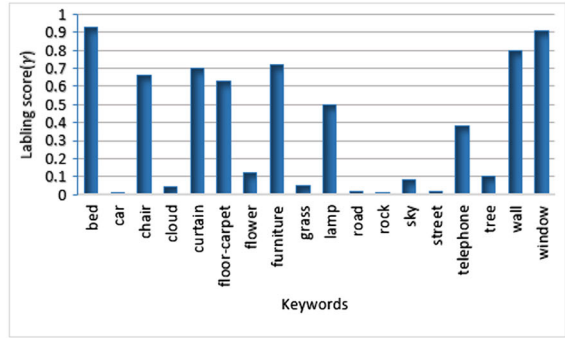
$$f(x) = \frac{1}{1 + e^{-\beta x}} \quad (4)$$

where  $\beta$  is a smoothing factor. This function generates values between 0 and 1. To decide whether a concept is present in the final labeling set, the following annotation scheme is designed using the sigmoid function as:

$$\gamma_m = f\left(\sum_{p=1}^T w_{m,p} \theta_p + b_m\right) \quad (5)$$

where  $\gamma_m$  denotes labeling score for the  $m$ -th concept for an image with  $T$  as the number of initial labels, and  $b_m$  stands for the number of times the  $m$ -th concept is repeated in the image. In this equation,  $\gamma_m$  is mainly determined by  $\sum_{p=1}^T w_{m,p} \theta_p$ , where  $w_{m,p}$  is the occurrence distribution of the  $m$ -th concept given the  $p$ -th concept in the initial label set. As shown in Fig. 3, the final annotations are generated probabilistically, which can also be useful in ranking the keywords. In spite of the existing methods that involve a limited number of neighbors as contextual constraints in analyzing the semantic context [56, 58], the correlations among all the candidate local labels are investigated in this work. For instance, to assign the label ‘curtain’ to an image, the existence of correlated labels such as window, and the repetition of this label are taken into account. Thereafter, the consistency of the set of semantic concepts is maintained according to the contextual relationships.

The model parameters are obtained by the training the model, when the correlative information of keywords is stored in matrix  $W^c$  [11]. This matrix analyzes the keyword



$$y_{bed} = f \left( \sum_{p=1}^{18} w_{bed,p} u_p + b_{bed} \right) = 0.93$$

**Fig. 3** An image and its labeling scores. The final annotations include ‘bed’, ‘window’, ‘wall’, ‘furniture’, ‘curtain’, ‘chair’, ‘floor-carpet’, ‘lamp’, ‘telephone’

correlations of training images. The correlation coefficients are normalized to  $[-1, 1]$ , where the negative values represent the inhibition of a specific concept by another in the same image, e.g., “sun” and “moon”. Our proposed method utilizes a nonlinear network, which is more efficient to represent complex nonlinear relationships among semantic concepts.

### 4.4 Complexity analysis

To annotate an untagged image, the trained model is used to probabilistically find the most relevant labels. In the training process, model parameters of different parts of the VSCIA, including clustering algorithm and scene-centric annotation are learned. However, in order to annotate images, the required computations in the learned model can be approximated as exploiting the most consistent descriptors to the given image, local labeling, and exploring the most relevant terms from the concept graph. We compare the computation complexity of the VSCIA with SML (supervised multiclass labeling) [9] which computes the conditional distributions of features based on the density estimation in the Gaussian mixture models. The SML method needs  $O(TLV)$ , where  $T$  denotes the number of visual descriptors in the given image,  $L$  stands for the number of Gaussian components, and  $V$  is the dictionary size [39]. On the other hand, the computation complexity of the proposed VSCIA to annotate an untagged image is  $O(VK)+O(TM\delta)+O(TM)$ , where  $K$ ,  $M$ , and  $\delta$  respectively denote the number of latent visual topics in pLSA, the number of semantic concepts, and the average number of sub-visual distributions in the mixture modeling. Obviously, it reduces to  $nmTM\delta$  in the real-world datasets, which is the dominant part of the required computations. Nevertheless, although the training time is considerable in the VSCIA, annotation time of the untagged images is much less than the that of SML since the number of sub-visual distributions is relatively small.

## 5 Experiment results

The proposed VSCIA approach is evaluated on three well-known multi-label benchmark databases, namely MSRC [46], IAPR TC-12 [15], and NUS-WIDE [12].

MSRC contains 591 images ( $240 \times 320$  pixels) associated with objects from 21 different semantic classes. The images are pixel-wise annotated, where each pixel is labeled with one



semantic class, or a ‘void’ class. One average, 3 labels are assigned to each image. From this dataset, 350 images are randomly selected for training and the remaining for testing.

The IAPR TC-12 dataset consists of a large set of semantic concepts and challenging images. It includes 20,000 pixel-wise annotated images, each of which annotated with a subset of labels from 275 semantic classes with an average of 5 labels per image. The collection provided is a subset of the MIR Flickr database [24] of real world images with varying lighting conditions, scales, positions and image qualities. This makes it a particularly challenging benchmark for automatic image annotation and retrieval. However, class labels are imbalanced, and some classes do not include enough images. Therefore, to accurately estimate the model parameters, these classes whose labels are assigned to less than 20 images are discarded. This results in a database of 19,970 images with 163 semantic labels, which is split into 17,970 training and 2000 test images.

To validate the performance of the VSCIA approach in large-scale real-world image environments, it is evaluated on the NUS-WIDE image dataset which contains 269,648 web images crawled from image sharing Flickr database. Remarkably, this dataset is one of the largest publicly available multi-label datasets with a wide range of classes from objects (e.g., plane and tree) to visual scenes (e.g., garden and temple). The dataset is associated with 81 ground truth semantic concept tags. It is separated into two parts, the first part contains 161,789 images for training and the second part contains 107,859 images for testing. The details of the NUS-WIDE dataset along with the two other datasets are summarized in Table 1.

The performance metrics used are precision, recall and F-measure. Given a concept  $w$ , let  $N_c$  be the number of images that are correctly annotated by  $w$ ,  $N_s$  be the total number of images that are annotated by  $w$ , and  $N_t$  be the number of images that include  $w$  in the ground-truth. The annotation precision and recall for the concept  $w$  are then defined as  $P(w) = \frac{N_c}{N_s}$ , and  $R(w) = \frac{N_c}{N_t}$ , respectively. To combine recall and precision into a single efficiency measure, we use the F-measure as  $F(w) = \frac{2 * P * R}{P + R}$ . The average precision, recall and F-measure over all concepts are used for evaluation.

Six state-of-the-art algorithms are used as benchmark baselines, namely SIRBOT (semantic image retrieval based on object translation) [62], HDIALR (hidden-concept driven image annotation and label ranking) [5], ECMRM (extended cross-media relevance model) [57], SML (supervised multiclass labeling) [9], PJMF (projective matrix factorization with unified embedding) [32], and HSIAHIT (high-level semantic image annotation based on hot Internet topics) [55]. The same parameter settings suggested by the authors are used for fair comparison. In SIRBOT, the JSEG tool is used to segment

**Table 1** A summary of the datasets used for experiment

Dataset	MSRC	IAPR TC-12	NUS-WIDE
Number of images	591	19,970	269,648
Image Size	240 × 320	480 × 360	varying sizes
Number of training images	350	17,970	161,789
Number of testing images	241	2000	107,859
Number of labels	21	163	81
Label per image	3	5	2
Image per label	50	350	6220

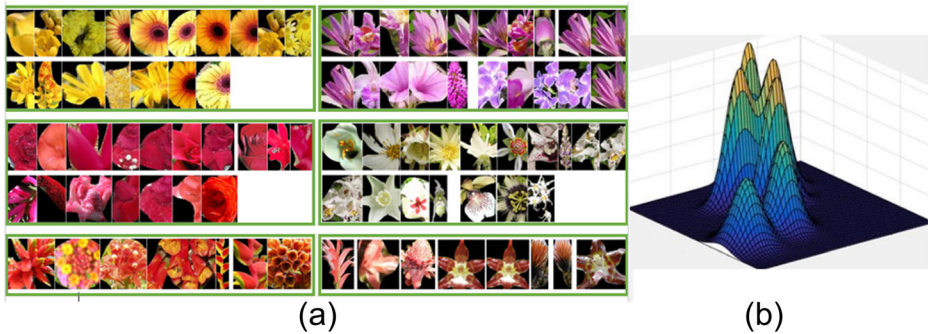
the dataset images into regions. The image regions are then represented by color, texture and shape features. In this method, a set of visual dictionaries are built by employing the adaptive vector quantization, each region is labelled with one semantic concept by a decision tree, and images are indexed and retrieved using an inverted file structure. HDIALR builds a relation matrix by decomposing the data matrix of the holistic image feature of the labeled and unlabeled images into two matrices of hidden concept basis matrix and hidden concept coefficient matrix. The relation matrix discovers the relationships between hidden concepts and labels. ECMRM incorporates regional, global, and contextual features to annotate images. The regional features are obtained from segmented regions, the global features are modeled as a global distribution of visual topics over an image, and the textual context is described by a multinomial distribution of keywords. In this model, both the global and contextual features are learned from the training data. In SML, the annotation problem is formulated as a supervised multiclass labeling. It is based on the density estimation to compute the conditional distributions of features given a certain keyword. To represent images in SML, the bags of localized features are utilized, and a Gaussian mixture model (GMM) consisting of 64 components is established. For the other two methods of PJMF, and HSIAHIT, we use the results listed in their original works. Accordingly, four methods of SIRBOT, HDIALR, ECMRM, and SML are employed for the performance comparison on the MSRC and IAPR TC-12 datasets. In addition, the VSCIA approach is compared with the PJMF and HSIAHIT methods on the full scale of the NUS-WIDE dataset for large-scale evaluation.

In the VSCIA method, a 192-dimensional feature vector is used to constitute the representation space. It includes the following robust features:

- i) color histograms from the RGB and LAB color spaces, with 16 bins in each color channel,
- ii) the mean and standard deviation of 512-dimensional Gist features,
- iii) first and second derivatives of Gaussian and Laplacian of Gaussian with various orientations and scales,
- iv) mean and standard deviation of Gabor texture features (at 5 scales and 8 orientations),
- v) seven invariant moments, and
- vi) three Tamura features (coarseness, contrast, directionality).

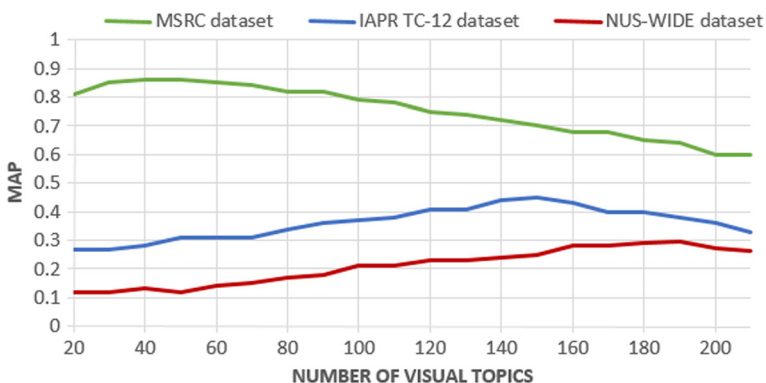
Since the ground-truth object annotations of the datasets are available, they are directly applied to the clustering algorithm to obtain sub-visual distributions. In the clustering algorithm,  $m_{pts}$  is empirically set to 3, which is sufficiently large to avoid wrongly clustered noise data, and small enough to cater for separating different object appearances into different clusters. For example, the samples for the semantic concept ‘flower’ in Fig. 4a are divided into 6 separate clusters according to their visual attributes. These clusters are characterized by different density levels. It is interesting to observe that some classes which indicate small visual diversity generate fewer clusters, and vice versa. For instance, semantic class ‘cloud’ is clustered into 2 clusters, whereas the visually complex classes such as ‘car’ generate more categories.

Once the sub-visual distributions are established, they are used to learn the mixture models for semantic classes. For each class, a mixture model is generated with expectation-maximization on the samples that are known to belong to the visual clusters of that class. Figure 4b illustrates the mixture model learned for the semantic concept ‘flower’ for the clusters shown in Fig. 4a.

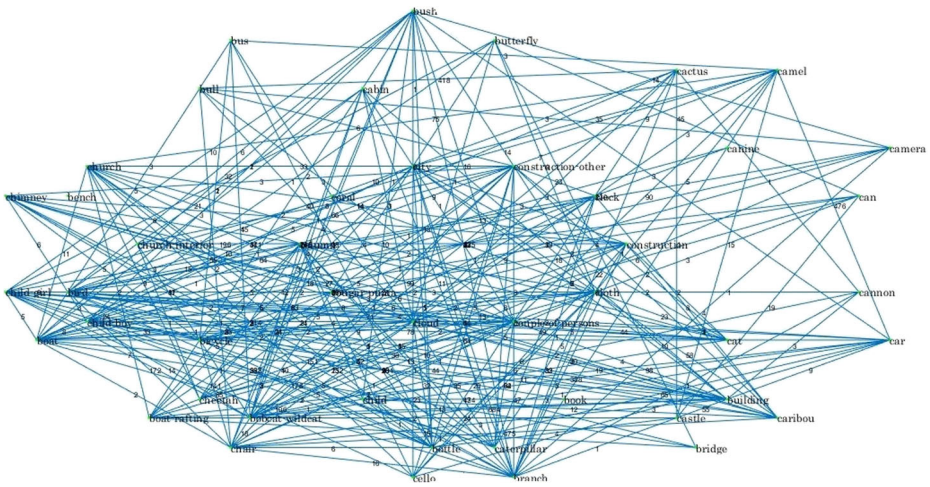


**Fig. 4** Semantic concept ‘flower’, **a**) distributed in 6 clusters by the clustering algorithm, and **b**) modeled by a mixture model learned from the samples of the clusters

The quantized visual descriptors are then used in a pLSA model to learn the visual topics. Obviously, a strong coherence is assumed to exist among visual attributes of an image. In our approach, the pLSA model assumes a latent lower dimensional topic space as the origin of the image scenes. In the generative process of pLSA, the joint distribution of the local descriptors and the images are used to model the visual topics. Once the visual topics are learned from the training data, each image can be viewed as representing one or more of these topics. Notably, each extracted local descriptor from a test image should be consistent with the other descriptors extracted from that image as they represent the same topics. The number of visual topics is set empirically by the cross-validation on the training set. Therefore, the mean average precision (MAP) is measured by varying the number of topics in each dataset. Specifically, by splitting the training set into 20 subsets, a single subset is used as the validation data to test the model, and the remaining subsets are utilized as training data. Then, this process is repeated by changing the number of visual topics from 20 to 210. Figure 5 depicts the MAP curves for different numbers of topics in the three datasets of the MSRC, IAPR TC-12, and NUS-WIDE. The MAP curve for MSRC peaks at 40 topics whereas for IAPR TC-12 peaks at 150 topics. On the other hand, the MAP curve for NUS-WIDE images shows its highest value at 190 topics. As shown in Fig. 5, the small numbers of topics lead to less satisfactory MAPs, which is due to the bias towards popular scene elements such as ‘sky’, ‘cloud’ and ‘water’. The MAPs increase with the number of topics growing, and peak at 40, 150, and 190. However, the MAPs decrease gradually when the number



**Fig. 5** The effect of the number of visual topics on the performance in different datasets



**Fig. 6** A part of the Concept graph constructed from co-occurrence matrix of concepts from IAPR TC-12 dataset

of topics is larger than the peak point. This is because the large number of topics leads to a sparse distribution of descriptors among the topics.

The latent visual topics can preserve the visual consistency among local descriptors of a specific image. With the consistent visual descriptors, the local labels are assigned to images using the learned mixture models. As images are also consistent in the annotation level, the assigned keywords should be consistent as well. Therefore, the semantic context is leveraged in this approach. The proposed concept graph is constructed from the co-occurrence matrix of keywords to exploit the semantic context. A part of the constructed graph for co-occurred concepts of IAPR TC-12 dataset is illustrated in Fig. 6. Each co-occurrence of two concepts increases their connection weight. For instance, it can be seen that ‘chair’ occurs together with ‘table’, ‘window’, ‘door’, ‘curtain’, and ‘dish’, 470, 252, 120, 107, and 94 times, respectively. This emphasizes the fact that images containing these objects show indoor scenes. Note that the average co-occurrence frequency is set for each concept based on the number of images in the database, which contain that concept.

Tables 2 and 3 summarize experimental results when comparing the proposed approach to the aforementioned state-of-the-art algorithms on the MSRC and IAPR TC-12 datasets, respectively. Note that all metrics are average bests, that is the best average to represent over the iterations in the experiments. When using the MSRC dataset, SIRBOT [62] performs better than SML [9], ECMRM [57] and HDIALR [5]. This is because ECMRM and HDIALR consider the label correlations, which is not completely applicable in the small label set of

**Table 2** Performance comparison of different approaches for image annotation on the MSRC dataset

Method	Precision	Recall	F-measure
SIRBOT [62]	0.72	0.74	0.73
ECMRM [57]	0.69	0.71	0.70
HDIALR [5]	0.63	0.75	0.69
SML [9]	0.70	0.72	0.71
VSCIA	0.77	0.82	0.79

**Table 3** Performance comparison of different approaches for image annotation on the IAPR TC-12 dataset

Method	Precision	Recall	F-measure
SIRBOT [62]	0.41	0.30	0.35
ECMRM [57]	0.28	0.34	0.31
HDIALR [5]	0.45	0.31	0.37
SML [9]	0.36	0.31	0.33
VSCIA	0.53	0.36	0.43







MSRC. The HDIALR tries to find correlation information among different labels by projecting the low-level image features onto a low-dimensional subspace. Instead, SIRBOT finds the shared information by computing the co-occurrence of objects in an image. However, though SIRBOT builds large visual dictionaries by separately performing vector quantization of local features, intra-class diversity makes the representative codewords to be converged inefficiently. Although SML is based on the sound application of multiple instance learning, it does not consider the multi-label relationships in image annotation. Moreover, its performance much drops if a concept includes a wide variability in its visual appearance. Remarkably, the VSCIA method achieves the highest F-measure of 0.79 for automatic image annotation in the MSRC dataset. The HDIALR algorithm outperforms SIRBOT, SML and ECMRM in IAPR TC-12 dataset as it discovers hidden visual concepts by constructing a relation matrix from the large number of holistic image features. Although the label correlations and intra-label diversity are taken into account, the visual and semantic context modeling cannot properly be specialized for the images. The highest F-measure of 0.43 is obtained by the VSCIA method in IAPR TC-12 dataset. As expected, the proposed VSCIA method provides greater performance in terms of average precision, recall and F-measure in both datasets since the consistency of both visual content and semantic context is ensured. In particular, when the data is sparse, it is important to capture the closeness information of all data samples, which can be well performed by considering the semantics of images consistent with the human visual system, thus achieving the best performance.

To provide the experimental comparison more convincing and feasible, the proposed VSCIA approach is evaluated on the large-scale NUS-WIDE image collection. We compare it with two recently developed algorithms of PJMF [32], and HSI AHIT [55]. In the PJMF, an unlabeled image is annotated by exploring its neighbors from feature and tag latent representations, which are embedded in a unified space using a transformation matrix. This matrix is an image-tag correlation matrix which is calculated using the original tagging information during the learning process. The HSI AHIT builds latent topics by analyzing the texts on the related web pages through a Latent Dirichlet Allocation (LDA) model. It tries to discover three kinds of relevant relationships between topics, topics and images, and images. In this method, images are annotated

**Table 4** Performance comparison of different approaches for image annotation on the NUS-WIDE dataset









Method	Precision	Recall	F-measure
PJMF [32]	0.20	0.12	0.15
HSI AHIT [55]	0.23	0.17	0.20
VSCIA	0.26	0.23	0.24



MSRC images						
Ground-truth	dog, grass, road, tree	building, flower person, road, sky	car, grass, road, sky, tree	boat, building, sky, tree, water,	building, person, road, sky	car, grass, road, sky, tree
SIRBOT	cat, dog, grass	bird, building, grass, person, sky, tree	building, car, sky, tree	boat, sky, tree, water	bicycle, building, sky, tree	car, person, road sky, tree
ECMRM	dog, grass, road, sheep, tree	boat, chair, person, sky, tree	car, dog, road, sky, tree	bird, boat, grass tree, water	aeroplane, building, person, sign, sky	boat, car, grass, sky, tree
HDIALR	cat, flower, road, tree	flower, person, sheep, sky, tree	bicycle, car, grass, road	boat, cow, grass, mountain, water	building, person, road, sky, tree	cow, grass, road, sky, tree
SML	dog, car, road, tree, water	aeroplane, flower, person, road, sky	car, grass, person, sky, tree	boat, car, road, tree, water	building, chair, person, road, sky	building, cow, grass, sky, tree
Predicted labels by VSCIA	dog, road, tree	flower, mountain, person, road, sky	car, road, sky tree	boat, building, sky, tree, water	building, person, road, sky	car, building, grass, road, sky, tree

**Fig. 7** Comparison of VSCIA annotations with those of ground-truth and the other methods in the MSRC dataset

by exploring the top keywords in the corresponding hot Internet topics which are formed by the modeling and clustering the built complex graph from the relevance relations. Table 4 reports the comparison results of the proposed VSCIA with these methods. It can be observed that the HSI AHIT method performs better than PJMF in terms of precision, recall, and F-measure. Although both PJMF and HSI AHIT methods consider the latent information in the image annotation, more relevant relationships are taken into account in the HSI AHIT method. In addition, the PJMF method adopts the Euclidean distance to assess the relevancy in the unified space, which is not reliable in the high dimensional image data. The local geometry information is

IAPR TC-12 images						
Ground-truth	cloud, eagle, sky, mountain, rock	flower, rock, sky	car, cloud, grass, house, sky, tree	cloud, grass, horse, house, man, mountain, tree	curtain, window, fabric, wall, bed, wooden furniture	bus, house, sky, street, tree
Predicted labels by VSCIA	cloud, eagle, sky, mountain	flower, sky	cloud, grass, house, sky, tree	cloud, grass, horse, house, mountain, tree	curtain, window, fabric, wall, bed, wooden furniture	bus, house, sky, street, tree
NUS-WIDE images						
Ground-truth	animal, dog, garden, toy	flower, garden, grass	car, cloud, sky	house, rainbow, road, sky	beach, cloud, ocean, sand, sunset	flower, water
Predicted labels by VSCIA	animal, dog, garden	flower, garden, grass, tree	car, cloud, sky	house, road, sky, sun	beach, cloud, ocean, sand	flower

**Fig. 8** Comparison of VSCIA annotations with those of ground-truth in the IAPR-TC 12, and NUS-WIDE datasets



also not effectively preserved in the visual and tag spaces. Despite the advantages of the HSI AHIT over the PJMF, the visual diversity of the semantic concepts, and the visual consistency of the image scenes are not investigated in this method. On the other hand, the VSCIA outperforms the annotation algorithms by achieving 0.26, 0.23, and 0.24 for precision, recall, and F-measure, respectively. This is because the VSCIA method works in a similar way to the human perception in explaining the image content. Therefore, the visual scene is perceived without ambiguity, and it is described by correlated and semantic terms. In other words, the VSCIA can better propagate the relationships among images and tags, and capture the consistencies between visual similarity and tag relevance.

Other examples of annotation results for different images from the MSRC dataset are shown in Fig. 7. The comparisons of the VSCIA annotations with the ground-truth in both IAPR-TC 12, and NUS-WIDE datasets are illustrated in Fig. 8. These examples suggest the effectiveness of VSCIA to annotate difficult images with multiple semantic concepts and visual diversity. It is noticeable that the number of training samples of some classes is unbalanced. This causes the wrong labels for some images. One of the key challenges in the image annotation task which is easily and automatically relaxed in the proposed model is the number of keywords that are necessary to describe the content of an image. Unlike many other algorithms that assign an exact number of labels to each image (usually 5 labels) [9, 38, 57], the images are annotated in the VSCIA by refining the local labels in the final label set.

## 6 Conclusion

A visual- and semantic-consistent image annotation (VSCIA) framework is presented in this paper. The intuition is that not only visual descriptors representing a scene should be correlated, but also annotation keywords must be consistent and specific for an image. This is similar to the human perception in explaining the image content, by first well-perceiving the visual scene without ambiguity, and then describing the image by correlated and semantic terms. Hence, the visual consistency, semantic independency, and their connections are taken into account. To this end, the pLSA model assumes each image is drawn from dependent visual topics, and a concept graph captures the semantic contextual correlations. Moreover, the intra-class diversity of the semantic concepts is already addressed by the clustering algorithm to increase the intra-class weights while reducing the inter-class similarities, and to boost the discriminative power of the local features. Besides, the potential intra-class samples in sub-visual distributions are linked by mixture models. Since this approach considers the coherency in two levels, it can construct less ambiguous local labels, and more realistic and meaningful correlated annotations with smaller semantic gap.

However, the direct connections of the visual topics and semantic concepts needs further exploration. It can lead to discovering semantic visual units underlying various forms of semantic expression. We investigate this point in our future work.

## References

1. Amiri SH, Jamzad M (2015) Automatic image annotation using semi-supervised generative modeling. *Pattern Recognit* 48(1):174–188
2. Ankerst M, Breunig MM, Kriegel H et al. (1999) “OPTICS: ordering points to identify the clustering structure.”. *Proc SIGMOD’99 ACM SIGMOD Int Conf Manag Data* 49–60

3. Assari SM, Zamir AR, Shah M et al. (2014) “Video classification using semantic concept co-occurrences.”. Proc 2014 I.E. Conf Comput Vision Pattern Recognit (CVPR) 2529–2536
4. Bannour H, Hudelot C (2014) Building and using fuzzy multimedia ontologies for semantic image annotation. *Multimed Tools Appl* 72(3):2107–2141
5. Bao B-K, Li T, Yan S (2012) Hidden-concept driven multilabel image annotation and label ranking. *IEEE Trans Multimed* 14(1):199–210
6. Barnard K, Duygulu P, Forsyth D, de Freitas N, Blei DM, Jordan MI (2003) Matching words and pictures. *J Mach Learn Res* 3:1107–1135
7. Bouveyron C, Girard S, Schmid C (2007) High-dimensional data clustering. *Comput Stat Data Anal* 52(1):502–519
8. Campello R, Moulavi D, Sander J et al. (2013) “Density-based clustering based on hierarchical density estimates.”. Proc 17th Pacific-Asia Conf Adv Knowledge Discovery Data Mining (PAKDD) Springer, LNAI 160–172
9. Carneiro G, Chan AB, Moreno PJ, Vasconcelos N (2007) Supervised learning of semantic classes for image annotation and retrieval. *Pattern Anal Mach Intell IEEE Trans* 29(3):394–410
10. Chapelle O, Haffner P, Vapnik VN (1999) Support vector machines for histogram-based image classification. *Neural Networks, IEEE Trans* 10(5):1055–1064
11. Chen Z, Fu H, Chi Z, Feng D (2012) An adaptive recognition model for image annotation. *Syst Man, Cybern Part C Appl Rev IEEE Trans* 42(6):1120–1127
12. Chua T, Tang J, Hong R et al. (2009) “NUS-WIDE: a real-world web image database from national university of Singapore.”. Proc ACM Int Conf Image Video Retrieval, Greece 48–53
13. Csurka G, Dance CR, Fan L et al. (2004) “Visual categorization with bags of keypoints.”. Proc Workshop Statistical Learn Comput vision (ECCV 2004) 1–22
14. Donoho DL, Grimes C (2003) Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proc Natl Acad Sci* 100(10):5591–5596
15. Escalante HJ, Hernández CA, Gonzalez JA, López-López A, Montes M, Morales EF, Enrique Sucar L, Villaseñor L, Grubinger M (2010) The segmented and annotated IAPR TC-12 benchmark. *Comput Vis Image Underst* 114(4):419–428
16. Ester M, Kriegel H, Sander J et al. (1996) “A density-based algorithm for discovering clusters in large spatial databases with noise.”. Proc Second Int Conf Knowledge Discovery Data Mining (KDD-96)
17. Fernando B, Fromont E, Tuytelaars T et al. (2012) “Effective use of frequent itemset mining for image classification.”. Proc Europe Conf Comput Vision (ECCV) 214–227
18. Gao Y, Ji R, Liu W, Dai Q, Hua G (2014) Weakly supervised visual dictionary learning by harnessing image attributes. *Imag Process IEEE Trans* 23(12):5400–5411
19. Gao S, Tsang IW, Chia L (2013) Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. *Pattern Anal Mach Intell IEEE Trans* 35(1):92–104
20. Gould S, Rodgers J, Cohen D, Elidan G, Koller D (2008) Multi-class segmentation with relative location prior. *Int J Comput Vis* 80(3):300–316
21. Guha S, Rastogi R, Shim K et al. (1998) “CURE: an efficient clustering algorithm for large databases.”. Proc SIGMOD’98 ACM SIGMOD Int Conf Manag Data 73–84
22. He K, Chan W, Zhu G, Lin L, Zhou X (2014) Image region labeling by exploring contextual information of visual spatial and semantic concepts. Proc 15th Pacific Rim Conf Adv Multimed Inform Process–PCM 1:93–102
23. He X, Zemel RS, Carreira-Perpinan MA (2004) “Multiscale conditional random fields for image labeling.”. Proc 2004 I.E. Comput Soc Conf Comput Vision Pattern Recognit (CVPR)
24. Huiskes M, Thomee B, Lew M et al. (2010) “New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative.”. Proc 11th ACM Int Conf Multimed Inform Retrieval 527–536
25. Izadinia H, Shah M (2012) Recognizing complex events using large margin joint low-level event model. Proc Europ Conf Comput Vision (ECCV) 7575:430–444
26. Jiang W, Chang SF, Loui AC (2007) Context-based concept fusion with boosted conditional random fields. Proc 2007 I.E. Int Conf Acoustics, Speech, Sign Process I(Pts 1-3):949–952
27. Ladick’ L, Russell C, Kohli P et al (2009) “Associative hierarchical crfs for object class image segmentation.”. Proc 12th IEEE Int Conf Comput Vision (ICCV) 739–746
28. Lazebnik S, Raginsky M (2009) Supervised learning of quantizer codebooks by information loss minimization. *Pattern Anal Mach Intell IEEE Trans* 31(7):1294–1309
29. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. Proc 2006 I.E. Comput Soc Conf Comput Vision Pattern Recognit (CVPR) 2:2169–2178
30. Leung CHC, Chan AWS, Milani A, Liu J, Li Y (2012) Intelligent social media indexing and sharing using an adaptive indexing search engine. *ACM Trans Intell Syst Technol* 3(3):1–27
31. Li L, Jiang S, Huang Q (2012) Learning hierarchical semantic description via mixed-norm regularization for image understanding. *IEEE Trans Multimed* 14(5):1401–1413
32. Li Z, Liu J, Tang J, Lu H (2014) Projective matrix factorization with unified embedding for social image tagging. *Comput Vis Image Underst* 124:71–78

33. Li J, Wang JZ (2008) Real-time computerized annotation of pictures. *IEEE Trans Pattern Anal Mach Intell* 30(6):985–1002
34. Liu W, Tao D (2013) Multiview hessian regularization for image annotation. *Imag Process IEEE Trans* 22(7): 2676–2687
35. Liu W, Tao D, Cheng J, Tang Y (2014) Multiview hessian discriminative sparse coding for image annotation. *Comput Vis Image Underst* 118:50–60
36. Llorente A, Manmatha R, Ruger S et al. (2010) “Image retrieval using Markov random fields and global image features,”. *Proc ACM Int Conf Imag Video Retriev* 243
37. Ma Z, Nie F, Yang Y, Uijlings JRR, Sebe N (2012) Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Trans Multimed* 14(4):1021–1030
38. Makadia A, Pavlovic V, Kumar S (2010) Baselines for image annotation. *Int J Comput Vis* 90(1):88–105
39. Nguyen C-T, Kaothanthong N, Tokuyama T, Phan X-H (2013) A feature-word-topic model for image annotation and retrieval. *ACM Trans Web* 7(3):1–24
40. Parsons L, Haque E, Liu H (2004) Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explor Newsl* 6(1):90–105
41. Rabinovich A, Vedaldi A, Galleguillos C et al. (2007) “Objects in context,”. *Proc 11th IEEE Int Conf Comput Vision (ICCV)* 1–8
42. Rasiwasia N, Vasconcelos N (2009) “Holistic context modeling using semantic co-occurrences,”. *Proc 2009 I.E. Conf Comput Vision Pattern Recognit (CVPR)* 1889–1895
43. Ritendra D, Joshi D, Li J, Wang JZ (2008) Image retrieval: ideas, influences, and trends of the new age. *ACM Comput Surv* 40(2):1–60
44. Russell BC, Efros AA, Sivic J, Freeman WT, Zisserman A (2006) Using multiple segmentations to discover objects and their extent in image collections. *Proc 2006 I.E. Comput Soc Conf Comput Vision Pattern Recognit (CVPR)* 2:1605–1614
45. Schultz M, Joachims T (2004) “Learning a distance metric from relative comparisons,”. *Adv Neural Inf Process Syst* 41–48
46. Shotton J, Winn J, Rother C et al. (2006) “Texonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation,”. *Proc Europ Conf Comput Vision (ECCV)*, Springer Berlin Heidelberg 1–15
47. Sivic J, Zisserman A (2003) Video Google: a text retrieval approach to object matching in videos. *Proc 9th IEEE Int Conf Comput Vision (ICCV)* 2:1470–1477
48. Song X, Jiang S, Wang S, Li L, Huang Q (2014) Polysemious visual representation based on feature aggregation for large scale image applications. *Multimed Tools Appl* 74(2):595–611
49. Su J-H, Huang W-J, Yu PS, Tseng VS (2011) Efficient relevance feedback for content-based image retrieval by mining user navigation patterns. *Knowl Data Eng IEEE Trans* 23(3):360–372
50. Tao D, Jin L, Liu W, Li X (2013) Hessian regularized support vector machines for mobile image annotation on the cloud. *Multimed, IEEE Trans* 15(4):833–844
51. Tirilly P, Claveau V, Gros P et al. (2008) “Language modeling for bag-of-visual word image categorization,”. *Proc CIVR’08 - Int Conf Content-Based Imag Video Retriev* 249–258
52. Vailaya A, Figueiredo MAT, Jain AK, Zhang H-J (2001) Image classification for content-based indexing. *Imag Process IEEE Trans* 10(1):117–130
53. Van Gemert JC, Snoek CGM, Veenman CJ, Smeulders AWM, Geusebroek J (2010) Comparing compact codebooks for visual categorization. *Comput Vis Image Underst* 114(4):450–462
54. Vogel J, Schiele B (2007) Semantic modeling of natural scenes for content-based image retrieval. *Int J Comput Vis* 72(2):133–157
55. Wang X, Du J, Wu S, Li X, Xin H, Zhang Y, Li F (2015) High-level semantic image annotation based on hot Internet topics. *Multimed Tools Appl* 74(6):2055–2084
56. Wang H, Lu T, Wang Y et al. (2014) “Weakly-supervised region annotation for understanding scene images,”. *Multimed Tools Appl* 1–25
57. Wang Y, Mei T, Gong S, Hua X-S (2009) Combining global, regional and contextual features for automatic image annotation. *Pattern Recognit* 42(2):259–266
58. Xiang Y, Zhou X, Liu Z et al. (2010) “Semantic context modeling with maximal margin Conditional Random Fields for automatic image annotation,”. *Proc 2010 I.E. Conf Comput Vision Pattern Recognit (CVPR)* 3368–3375
59. Xu C, Tao D, Xu C (2014) Large-margin multi-view information bottleneck. *Pattern Anal Mach Intell IEEE Trans* 36(8):1559–1572
60. Zha Z.-J, Hua X.-S, Mei T et al. (2008) “Joint multi-label multi-instance learning for image classification,”. *Proc 2008 I.E. Conf Comput Vision Pattern Recognit (CVPR)* 1–8
61. Zhang D, Islam MM, Lu G (2012) A review on automatic image annotation techniques. *Pattern Recognit* 45(1):346–362

62. Zhang D, Monirul Islam M, Lu G (2013) Structural image retrieval using automatic image annotation and region based inverted file. *J Vis Commun Image Represent* 24(7):1087–1098
63. Zhang S, Tian Q, Hua G, Huang Q, Gao W (2014) ObjectPatchNet: towards scalable and semantic image annotation and retrieval. *Comput Vis Image Underst* 118:16–29
64. Zhao S, Yao H, Yang Y et al. (2014) “Affective image retrieval via multi-graph learning.”. *Proc ACM Int Conf Multimed* 1025–1028
65. Zheng Y-T, Neo S-Y, Chua T-S, Tian Q (2009) Toward a higher-level visual representation for object-based image retrieval. *Vis Comput* 25(1):13–23



**Mohsen Zand** received his PhD degree in Multimedia Information Retrieval from the Universiti Putra Malaysia (UPM) in 2015. In 2005, he joined the department of Computer Engineering at the Islamic Azad University, where he is currently a full time faculty member. He was engaged in different projects such as medical imaging, and image processing. His research interests include multimedia content analysis, indexing and retrieval, pattern recognition, and machine learning.



**Shyamala Doraisamy** is an Associate Professor at the Department of Multimedia, Faculty of Computer Science and Information Technology, University Putra Malaysia (UPM). She received her PhD from Imperial College London in 2004, specializing in the field of Music Information Retrieval and currently heads the Digital Information Computation and Retrieval research group at UPM. Her research interest includes Multimedia Information Processing, focusing in particular on Audio Content Analysis and Applications for Music and Health Informatics. She is a committee member of the Malaysia Society of Information Retrieval and Knowledge Management (PECAMP) and the recent Tenth Asia Information Retrieval Societies (AIRS) conference.



**Alfian Abdul Halin** obtained his Master of Multimedia Computing from Monash University, Australia in 2004 and PhD in Computer Science (specializing in Sports Video Content Analysis) from Universiti Sains Malaysia, Penang in 2011. He joined the Faculty of Computer Science and Information Technology (FCSIT), Universiti Putra Malaysia in 2001 and now holds the position of senior lecturer. His research interests include semantic image/video retrieval using pattern recognition and machine learning.



**Mas Rina Mustaffa** received her BCompSc degree (Multimedia) and MSc degree (Multimedia Systems) in 2003 and 2006 respectively. She received her PhD for studies in Content-based Image Retrieval (CBIR) in 2012. All of the three degrees are from Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia. Her primary research interests are multimedia systems and applications, CBIR, image processing, pattern recognition, and interactive multimedia.

She is currently a Senior Lecturer at the Department of Multimedia, Faculty of Computer Science and Information Technology, UPM. She has authored several publications in various journals and proceedings and presented at many conferences. She also has been actively involved in several international conferences as technical program committee.

Dr. Mas Rina is a member of the ACM, IACSIT, IEEE, and PECAMP.