Visualizing Bag-of-Features Image Categorization Using Anchored Maps

Gao Yi [*]
The University of Tokyo
Tokyo, 113-8656 Japan

Hsiang-Yun Wu [†]
The University of Tokyo
Tokyo, 113-8565 Japan

Kazuo Misue [‡]
University of Tsukuba
Tsukuba, 305-8577 Japan

Kazuyo Mizuno [§]
The University of Tokyo
Tokyo, 113-8565 Japan

Shigeo Takahashi [¶]
The University of Tokyo
Tokyo, 113-8565 Japan

ABSTRACT

The bag-of-features models is one of the most popular and promising approaches for extracting the underlying semantics from image databases. However, the associated image categorization based on machine learning techniques may not convince us of its validity since we cannot visually verify how the images have been classified in the high-dimensional image feature space. This paper aims at visually rearrange the images in the projected feature space by taking advantage of a set of representative features called visual words obtained using the bag-of-features model. Our main idea is to associate each image with a specific number of visual words to compose a bipartite graph, and then lay out the overall images using anchored map representation in which the ordering of anchor nodes is optimized through a genetic algorithm. For handling relatively large image datasets, we adaptively merge a pair of most similar images one by one to conduct the hierarchical clustering through the similarity measure based on the weighted Jaccard coefficient. Voronoi partitioning has been also incorporated into our approach so that we can visually identify the image categorization based on support vector machine. Experimental results are finally presented to demonstrate that our visualization framework can effectively elucidate the underlying relationships between images and visual words through the anchored map representation.

[*]gaoyi@visual.k.u-tokyo.ac.jp
[†]hsiang.yun.wu@computer.org
[‡]misue@cs.tsukuba.ac.jp
[§]mizunok@visual.k.u-tokyo.ac.jp
[¶]takahashis@computer.org

VINCI 2014, August 5–8, 2014, Sydney, Australia.

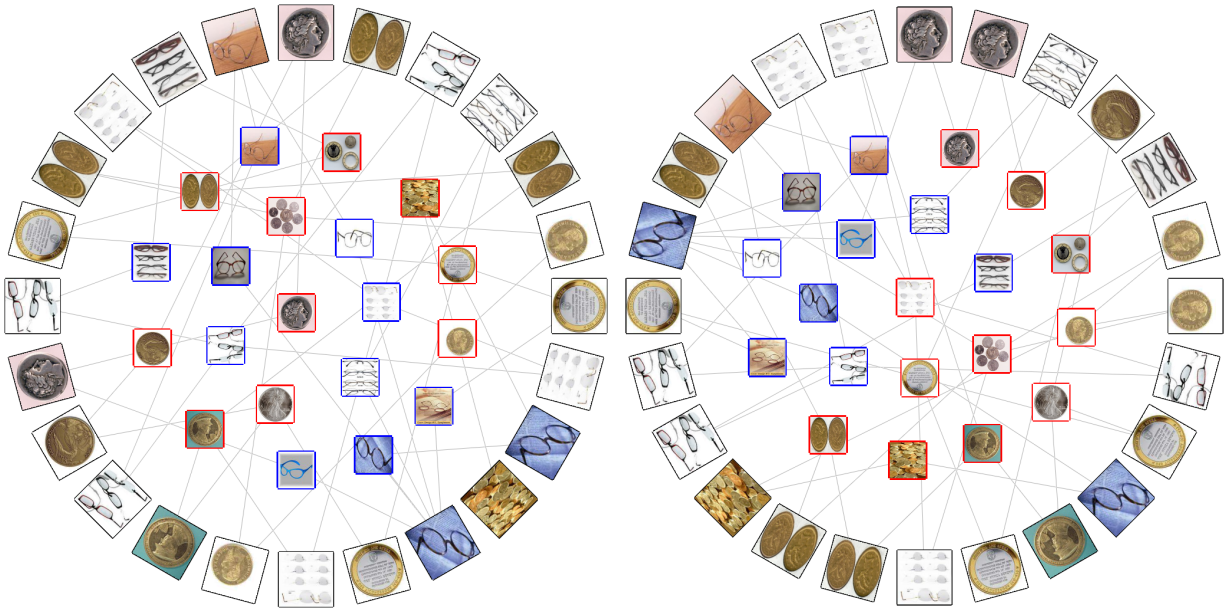
Author Keywords

bag-of-features, bipartite graphs, anchored maps, hierarchical clustering, weighted Jaccard similarity

INTRODUCTION

Sophisticating tools for image categorization becomes more crucial in content-based retrieval of image databases due to the rapid increase in their data sizes. While the associated techniques have been improved until recently, it is still labor intensive to sufficiently infer the underlying semantics from images. This problem primarily arises from the fact that we cannot precisely identify specific objects embedded in the images regardless of possible variations in their view, lighting, and occlusion conditions. The *bag-of-features* (BoF) model [19, 3] successfully alleviates this problem for effective image retrieval. A main idea behind the BoF model is to seek an analogy of methods for inferring text categorization based on the bag-of-words model, where each document is represented as a sparse vector of representative words by referring to their occurrence without worrying about their associated orders. In practice, the BoF model allows us to associate an individual image with a small weighted set of *visual words*, each of which stands for a group of local features in the high-dimensional feature space and thus corresponds to some specific image content in the image.

Nonetheless, the correctness of the image categorization is not always convincing even with the help of classification methods based on machine learning algorithms, since the actual mechanism for the associated image categorization has not been fully visualized due to the high-dimensionality of the image feature space. In this study, we solve this problem by encoding the relationship between images and visual words as a bipartite graph first, and then employing *anchored map* representation [13] to rearrange the image set on the 2D screen space, as shown in Fig. 1(a). Genetic algorithms have



(a) (b)

Figure 1. Using anchored maps to visualize bag-of-features image categorization. (a) Original layout. (b) Enhanced layout with an optimized circular ordering of visual words annotated with representative images. Images in same category are brought closer to each other. (#{visual words} = 24.)

also been employed to optimize the circular ordering of visual words around the image feature space, so that we can visually elucidate the underlying relationship between images and visual words, as shown in Fig. 1(b). Furthermore, we introduced hierarchical representation of the images by adaptively merging images according to their similarity values for effectively handing a large set of images. This hierarchical representation also facilitates users conduct the image categorization according to their preference by interactively selecting a training set of images for marching learning techniques.

The remainder of this paper is organized as follows: Sect. provides a study on conventional techniques for bag-of-features and bipartite graph visualization. Sect. describes how we can extract image features and construct the dictionary of visual words by extracting low-level image features. Sect. presents our approach to transforming the high-dimensional image feature space to an anchored map representation by referring to the bipartite relationships between images and visual words. After having presenting several experimental results to demonstrate the feasibility of our prototype system in Sect. , we conclude this paper and refers to future work in Sect. .

RELATED WORK

Content-based image retrieval has been a hot topic in the research on image processing, computer graphics, and multimedia. For effective search for specific contents, it is important to classify images into several categories by inferring semantics of visual features embedded in them. The *bag-of-features* (BoF) model is a well-known approach for such image representation and helps us categorize images by computing the number of occurrence of particular visual features contained in each image [19, 3]. This idea originates from the concept of *bag-of-words* that naturally allows us to classify documents by counting the number of particular words defined in the dictionary [8]. Indeed, this concept has been extended to the image databases where a set of local features called *visual words* is employed as the dictionary for the analysis of image contents.

In the early stage of approaches of this type, several studies focused on detecting global image features for encoding the image as a whole. Nonetheless, these features appeared to be inappropriate for the purpose of categorizing images because they are too sensitive to image transformations including translation, scaling, and rotation together with lighting conditions and occlusions. Lowe presented an feature detection technique called *scale-invariant feature transform* (SIFT) [11], which extracts local image features in a way that they are robust enough to the prescribed conditions. In practice, the visual words were obtained by collecting the SIFT features from a set of training images and employing the conventional k-means clustering to identify the corresponding cluster centers as the visual words.

As for practical approaches for image categorization, we first encode each image in the database as a weighted sum of the relevant visual words. Indeed, the BoF model facilitates us to assign a sparse vector representation of visual words to each image by quantizing it in terms of its associated visual words [19, 3] or L_1-norm

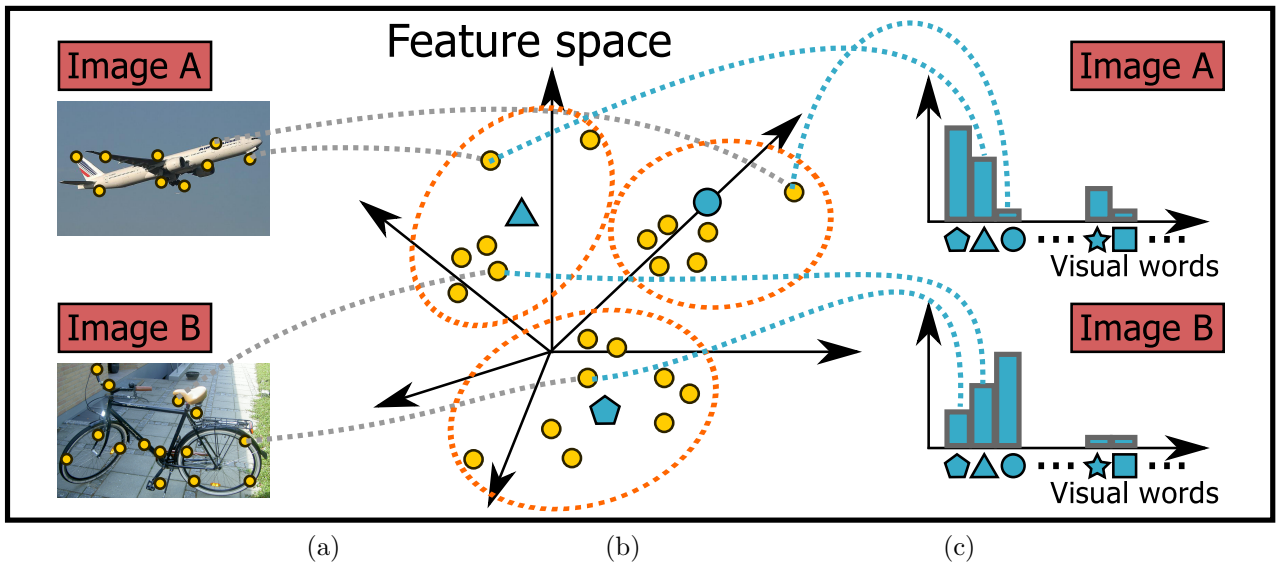


Figure 2. The BoF model. (a) SIFT feature vectors extracted from images are plotted in the 128 dimensional feature space. (b) The k-means clustering algorithm is employed to identify a visual word as the center of each cluster. (c) Each image is encoded as normalized histogram coordinates in terms of the visual words.

regularization [22]. *Support vector machine* (SVM) has been often employed as a standard classifier since it produces high accuracy in image categorization [3, 4]. As an extension, Bosch et al. [1] revisited the recognition scheme and apply it to the video by employing probabilistic latent semantic analysis (pLSA) followed by k-nearest neighbor (k-NN) classification. Over the years, a wide range of methods have been developed to improve the quality of the image categorization. A state-of-the-art technique is spatial pyramid matching (SPM) proposed by Lazebnik et al. [10], where they incorporated spatial gradient information of images at multiple scales into the BoF model. More studies also focused on improving the discriminative power of the visual words dictionary. For example, Winn et al. [21] introduced a statistical measure for the optimization framework to make the dictionary of the visual words more compact, while Perronnin [17] combined local and global feature detection frameworks to exhibit higher performance. However, the space of image features extracted by these approaches is always high-dimensional and too abstract to understand the meaningful structures hidden behind that space.

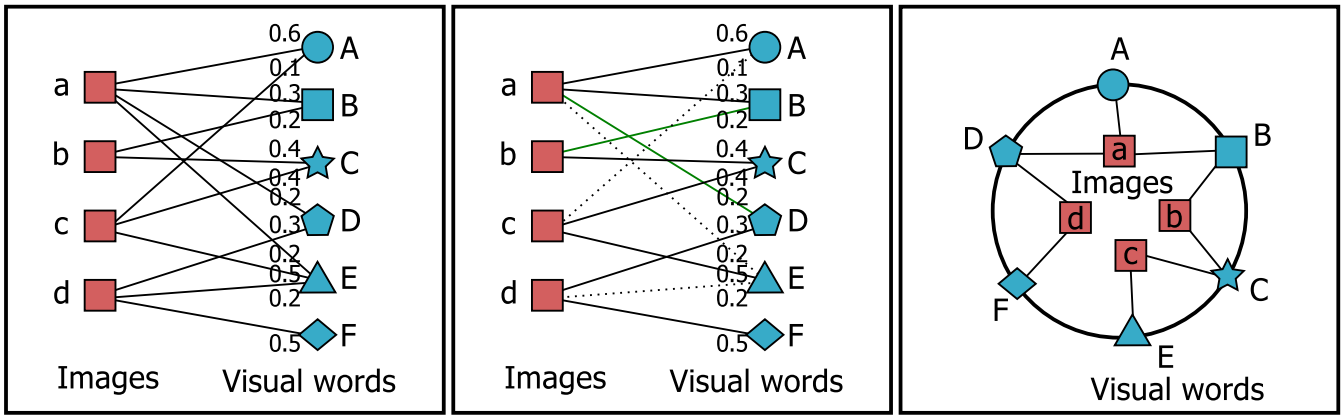
Visualizing high-dimensional feature space often successfully elucidates the image classification obtained through machine learning techniques. A dimensionality reduction technique called *multidimensional scaling* (MDS) [20, 9] is one of the common techniques to project the high-dimensional space onto a 2D screen space for better readability. Recently, Paulovich et al. [16] and Mamani et al. [12] developed dimensionality reduction frameworks that allows us to interactively edit the underlying structures of the high-dimensional space through screen-space manipulations. Furthermore, Mizuno et al. [15] presented a framework for interactively exploring feature space that is specific to the BoF models, by referring to the relationships between images and visual words. In our approach, we also focus on the such relationships specific to the BoF models and encode them as anchored map representations [13, 18] for visualization purposes. Technical details of the present approach will be detailed later in Sect. .

BAG-OF-FEATURES MODEL FOR IMAGE CATEGORIZATION

This section first provides a brief overview of the BoF model for encoding images as feature vectors, and describes how images are categorized using machine learning techniques.

Image Representation Based on the Bag-of-Features Model

In general, the BoF model consists of the three steps: *feature extraction*, *visual words dictionary formation* and *image-histogram representation*. The first step of the BoF construction is the *feature extraction*, where we extract SIFT features from the respective images. Here, the SIFT features are described as 128-dimensional feature vectors and plotted within the 128-dimensional feature space as shown in Fig. 2(a). For conducting the second step for the *visual words dictionary formation*, all the SIFT features are grouped into a specific number of clusters. The simplest technique for this purpose is the conventional k-means clustering algorithm, where the number of clusters k is predefined. Now we are ready to identify the center of each cluster as a representative feature called a *visual word*, and compose the list of k visual words as the dictionary as exhibited in Fig. 2(b). Our last step is *image histogram representation*, where we encode each image as a histogram coordinates in terms of the visual words. This is accomplished by quantizing each SIFT feature vector contained in the



(a) (b) (c)

Figure 3. Bipartite relationships between images and visual words in the BoF model. (a) An original bipartite graph. (b) A sparse bipartite graph after edge pruning. (c) The corresponding anchored map representation.

image to its closest visual word in the 128-dimensional feature space first, and then counting the occurrence of each visual word to construct the histogram. Finally, each image is represented as a sparse vector of visual words by normalizing the bins of the histogram to compose the normalized histogram coordinates, as shown in Fig. 2(c).

Image Categorization Using Support Vector Machine

In the BoF model, the support vector machine (SVM) is employed as the simplest learning models for classifying images by partitioning the high-dimensional space spanned by the extracted visual words [3]. In practice, the classifier finds the maximum marginal hypersurfaces that separates positive and negative samples in the training dataset, and further classify each of the unknown samples by referring to the separating hypersurfaces. In this paper, we introduce the SVM-based image categorization process proposed by Csurka et al. [3] and visualize how the bounding hypersurfaces enclose the images of specific type according to the input training samples provided by users. In our approach, we employed radial basis functions (RBFs) kernels for representing such separating hyperplanes to better classify the complicated configuration of images in the high-dimensional space, and visualize the associated image classification in the screen space for more convincing representation.

HIERARCHICAL BIPARTITE GRAPH VISUALIZATION

In this section, we describe how to visualize image categorization via an anchored map representation by referring to the bipartite relationships between images and visual words. We also introduce the weighted Jaccard similarity index for adaptively clustering images so that we can hierarchically represent large scale image sets within the framework of anchored maps.

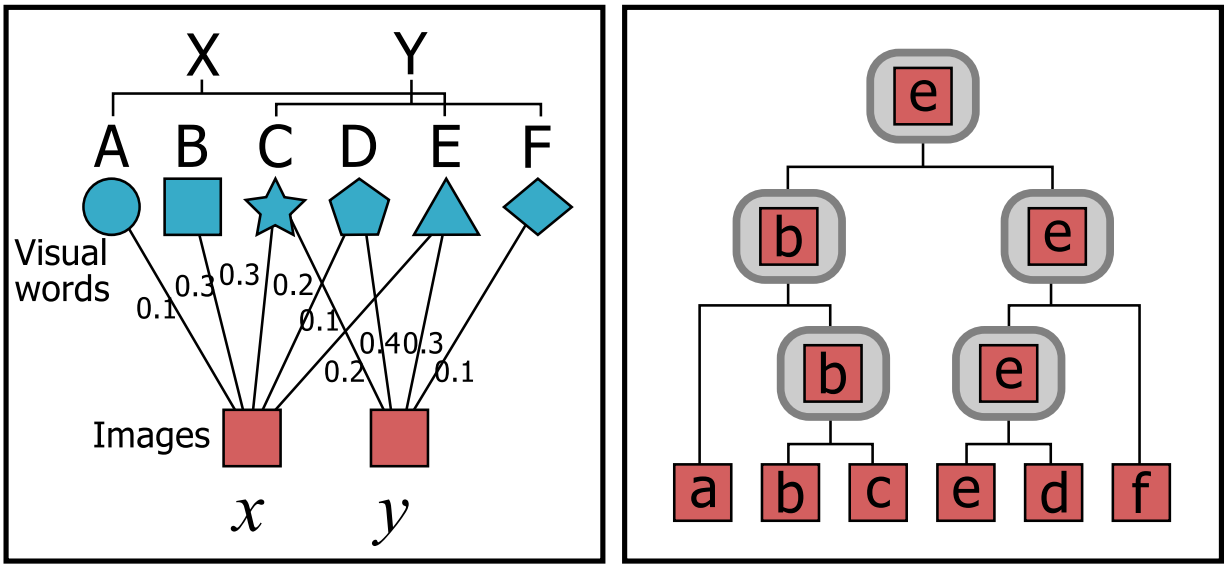
Bipartite Network Composition

The most common way of visualizing the high-dimensional image feature space is to employ dimensionality reduction techniques. Nonetheless, it is often the case that we still cannot fully discriminate each image category from others if the images are simply projected onto the low-dimensional space. Our original idea for alleviating this problem is to extract bipartite relationships between images and visual words from the BoF model first, and then transform them into a network structure so that we can take advantage of existing graph drawing techniques for better visualization.

For this purpose, we first establish edge connections between each image and its relevant visual words if they correspond to non-zero histogram coordinates of that image. Note that here we represent images and visual words as nodes of the bipartite graph, while we associate each normalized histogram coordinate value with the corresponding edge as its weight value as shown in Fig. 3(a). Furthermore, we would like to make the bipartite graph as sparse as possible for better readability of the resulting graph visualization. Thus, we sort the edges in an ascending order according to the weight values, and prune the edge having the minimum weight one by one until we cannot remove edges any more without decomposing the graph into multiple connected components, as shown in Fig. 3(b). In this way, we construct a sparse representation of the bipartite graph over the image and visual word nodes.

Anchored Map Representation

As for the visualization of the bipartite relationships, we employ anchored map representations formulated by Misue [13, 14]. In the anchored map representation, nodes in one of the two disjoint sets of the bipartite graph are equally spaced along the boundary of a disk region, while nodes of the other set are free to move within the disk, as shown in Fig. 3(c). In the figure, we release the image nodes within the central disk region of the anchored map and fixed the visual word nodes



(a) (b)

Figure 4. Hierarchical structure of bipartite graph visualization. (a) An example bipartite graph between images and visual words. (b) Dendrogram-based representation of clustered images.

along its circular boundary. The conventional spring embedder algorithm is also applied to the free nodes to avoid unnecessary overlap among images in the central region, where we also incorporate edge weights into our formulation so that each image will be brought closer to its relevant visual words according to their corresponding normalized histogram coordinates.

In our sparse representation of the bipartite graph, each image usually depends on a small number of visual words. This means that our scheme is more likely to bring image of the same category close to each other in the anchored map representation since they usually share almost the same set of visual words in their histogram representation. Furthermore, this visual readability of the image categorization can be enhanced if we carefully reorder the visual word nodes along the circular boundary of the disk to make each image node have its neighbor visual word nodes within its vicinity. This is accomplished by devising genetic-based algorithms for optimizing the circular ordering of visual words, where we define a chromosome as a value-encoding sequence of visual word IDs. For fully discriminating between image categories, we optimize the chromosome sequence by defining the cost function so that, for each image node, every pair of its adjacent visual word nodes become closer to each other. This amounts to calculating the circular distance between adjacent visual word nodes for each image node, and summing up the squared distances except for the largest one [13]. This genetic-based optimization provides us with better anchored maps in the sense that images in the same category will be closer to each other in the central disk region as shown in Fig. 1(b).

Hierarchical Clustering of Images

As the number of input images increases, the central disk region of the anchored map will be more crowded with the images. For improving the scalability of the anchored map representation, we also introduced hierarchical representation of the anchored map by adaptively clustering images according to their image similarities. More specifically, we compose a dendrogram tree structure of images by merging a pair of the most similar images one by one iteratively [18]. For evaluating the similarity among images, we employ the conventional Jaccard similarity index, which is the most popular similarity measure between a pair of sets [2]. Let us consider two sets X and Y for example. The conventional Jaccard index is defined as $J(X,Y) = |X \cap Y|/|X \cup Y|$, where $|Z|$ represents the number of elements contained in the set Z. However, in our case, the weighted Jaccard similarity index [7, 2] is more appropriate in the sense that we can incorporate the importance of each relevant visual word when evaluating the image similarities, rather than simply counting the number of relevant visual words in the union and intersection of the two sets.

As described earlier, our bipartite graph is composed by connecting an image with its relevant visual words, and the weight of each edge is equivalent to the normalized histogram coordinate value of the corresponding visual word with respect to that image. Thus we can easily compute the weighted Jaccard similarity index between a pair of images by referring to their corresponding sets of visual words X and Y, together with their corresponding edge weights, as follows:

$$\mathrm{WJ}(X,Y) = \frac{\sum_{i=1}^{n} \min(X_i, Y_i)}{\sum_{i=1}^{n} \max(X_i, Y_i)} \quad (1)$$

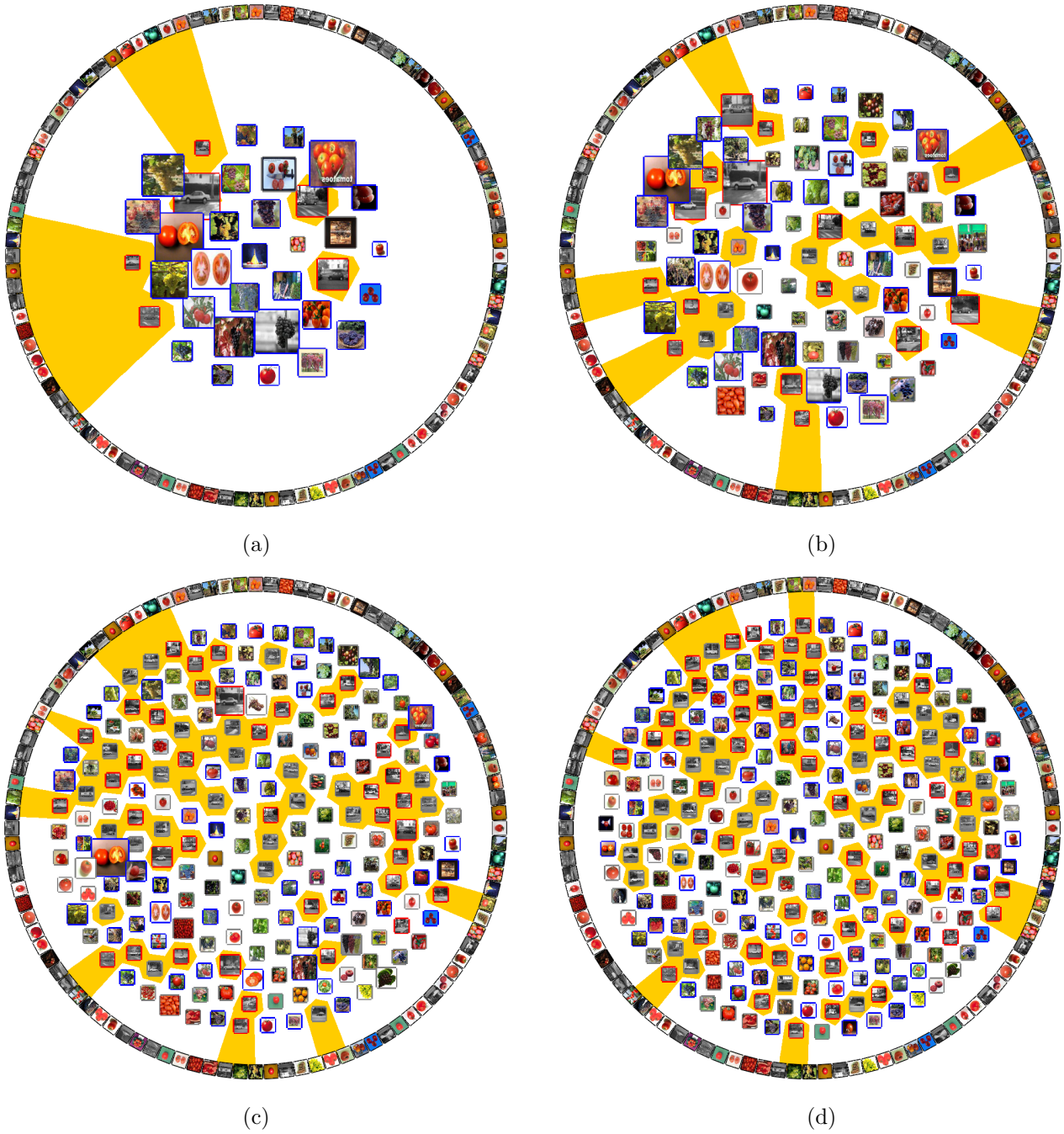


Figure 5. Discriminating car images using the support vector machine at multiple hierarchical levels. Images of the training set are labeled as red (car images) and blue (others). The inferred region of the car images is rendered in yellow through the Voronoi tessellation. (a) 10%, (b) 30%, (c) 40%, and (d) 100% of images. (#{visual words} = 100.)

where n denotes the total number of visual words contained in the union of X and Y. Note that the numerator is obtained by summing up the minimum values between two weights of the edges emanating from visual words in X and Y, while the denominator is the sum of the maximum values. Fig. 4(a) shows an example, where the X_i and Y_i are defined as normalized histogram coordinates for the image nodes x and y, and thus we can set $(X_i) = (0.1, 0.3, 0.3, 0.2, 0.1, 0.0)$ and $(Y_i) = (0.0, 0.0, 0.2, 0.4, 0.3, 0.1)$. This means that we can compute the weighted Jaccard similarity index between the image nodes x and y as

$$\mathrm{WJ}(X, Y) = \frac{0.0 + 0.0 + 0.2 + 0.2 + 0.1 + 0.0}{0.1 + 0.3 + 0.3 + 0.4 + 0.3 + 0.1} = \frac{1}{3}.$$

Using the weighted Jaccard measure, we can iteratively merge a pair of the most similar images into a group one by one, and encode the clustering process as a den-

drogram tree representation as shown in Fig. 4(b). As illustrated in this figure, we incorporate an image node having a smaller number of child nodes into the other image node representing more child nodes in our implementation.

Visualizing SVM-based Image Classification

We also equip our prototype system with an interface for classifying images using support vector marching (SVM). In practice, users are allowed to interactively specify a subset of images as a training set for SVM-based classifier together with the tags that represent whether the corresponding images are classified into a specific category or not. Nonetheless, conventional BoF models just present the classification results only and do not provide us with any information about how the images are classified in the high-dimensional image feature space. When projecting the high-dimensional image categorization onto the central disk region within the anchored map, we introduced the Voronoi tessellation technique in order to clarify how the region is partitioned according to the image categorization. Here, we employ the position of each image node as a seed point for the Voronoi cell, and assign a specific color to that cell according to its image category obtained through the SVM classification. This successfully makes us convinced with the image categorization provided by the SVM-based classifier by visualizing the associated image categorization within the anchored map representation. Note that, in our implementation, we incorporated a hardware-assisted algorithm for computing Voronoi diagrams [6] and restrict the drawing area to the central disk region of the anchored map using the stencil buffer.

RESULTS

Our system has been implemented on a laptop PC with an Intel Core i7 CPU (2GHz, 4MB cache) and 8GB RAM, and the source code has been written in C++ using the OpenGL library for drawing graph layouts, OpenCV library for SIFT feature extraction and SVM learning models, and GAlib library for the implementation of the genetic-based algorithm. The images datasets used in this paper were collected from Caltech256 [5].

Fig. 1 exemplifies how the underlying image categorization can be better visualized by taking advantage of the optimal ordering of visual words around the circular boundary of the anchored map representation. Here, Fig. 1(a) shows the initial ordering of visual words and layout of images in the dataset where images of coins and spectacles are intricately mixed. On the other hand, images of two categories are sufficiently discriminated in Fig. 1(b) when we rearrange the ordering of the visual words using genetic-based optimization. The image set exhibited in Fig. 5 contains images of three different objects, i.e., cars, tomatos and grapes from which we try to discriminate car images specifically from the others. For effectively handling a large number of images, we first compute a small number of image clusters through hierarchical grouping of images, and distinguish car images from the others as our target using the SVM-based image categorization. Note that here the images outlined in red are labeled as example images within the specific category (i.e. car images), while those in blue are images that are out of our target. We then gradually decompose each image cluster into smaller clusters, and adjust the image categorization by interactively labeling a small number of images as the training set according to their categories. This successfully allows us to enclose car images within yellow background region from the coarsest level to the finest (i.e. original) level as shown in Fig. 5. Fig. 6 demonstrates how we can categorize images of a specific category even when we train our image classifier indirectly with similar looking images. In this case, we represent each image in terms of visual words obtained from training images containing tomatos, coins, and cars and try to collect images of round shapes. However, we also take as input images of additional categories such as CDs and glasses in this example, while we still can categorize images of round objects into our target category using the SVM-based classifier, and clearly visualize the associated image categorization both at coarse and fine levels through the anchored map representation as shown in the figure.

CONCLUSION

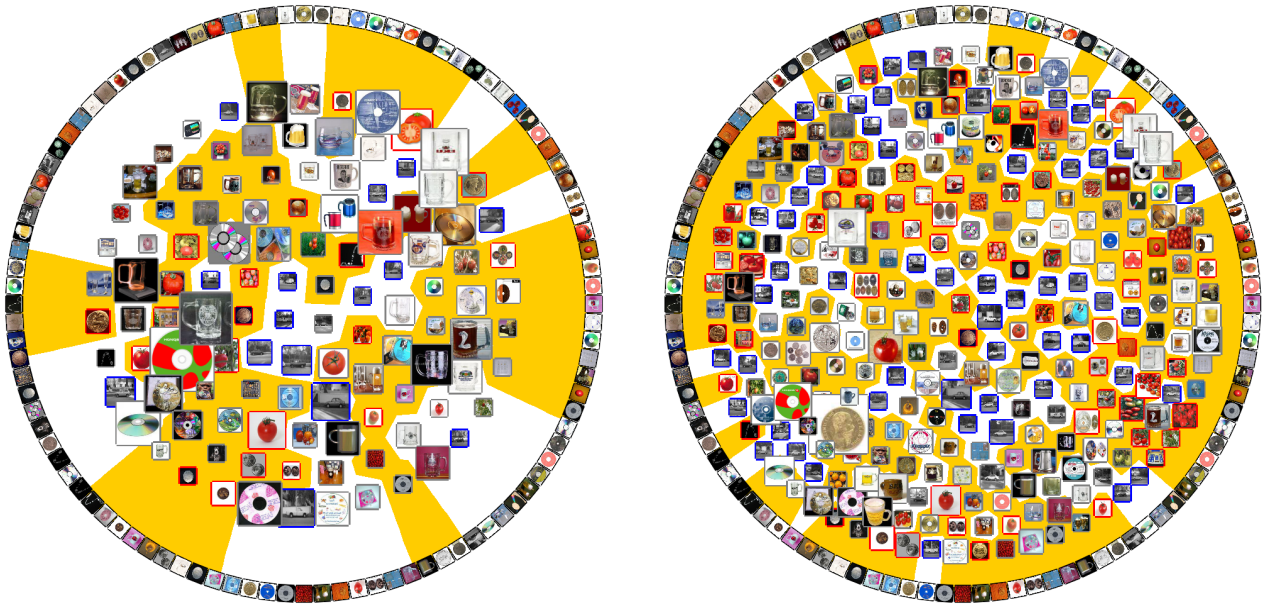
In this paper, we have presented an approach to visualizing image categorization within the high-dimensional feature space by taking advantage of the characteristics of the BoF model. The idea behind our approach is to extract the bipartite relationships between the input images and visual words first and then visualize them as a network using the anchored map representation. This new type of dimensionality reduction framework successfully convinces us of the plausibility of resulting image categorization based on the BoF model. The readability of the anchored map representations have been further enhanced by seeking the optimal circular ordering of visual words and dendrogram-based hierarchical representation of images. Voronoi-based partitioning has been also incorporated into the central disk region of the anchored map to visualize the border of some specific image category.

Fully classifying images of multiple categories according to users' preference remains to be tackled. The readability of the anchored map representations also depends on the quality of the sparse vector representations of the images in terms of the extracted visual words. Enhancing the interactivity of the present image retrieval system is also left as a future research theme.

ADDITIONAL AUTHORS

REFERENCES

1. Bosch, A., Zisserman, A., and Muñoz, X. Scene classification via pLSA. In *Proc. 9th European Conference on Computer Vision (ECCV 2006)* (2006), vol. 3954 of *Springer Lecture Notes in Computer Science*, pp. 517–530.



(a) (b)

Figure 6. Categorizing images of round objects from images of five different categories (tomatos, coins, cars, CDs, and glasses). (a) Coarse level. (b) Fine level. (#{visual words} = 100.)

2. Chierichetti, F., Kumar, R., Pandey, S., and Vassilvitskii, S. Finding the jaccard median. In *Proc. 21st Annual ACM-SIAM Symposium on Discrete Algorithms* (2010), pp. 293–311.

3. Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. Visual categorization with bags of keypoints. In *ECCV'04 Workshop on Statistical Learning in Computer Vision* (2004), pp. 1–22.

4. Eichhorn, J., and Chapelle, O. Object categorization with SVM: Kernels for local features. In *Advances in Neural Information Processing Systems (NIPS)* (2004).

5. Griffin, G., Holub, A. D., and Perona, P. The caltech 256. Tech. rep., California Institute of Technology, 2006.

6. Hoff III, K. E., Culver, T., Keyser, J., Lin, M., and Manocha, D. Fast computation of generalized voronoi diagrams using graphics hardware. In *Proc. SIGGRAPH '99* (1999), pp. 277–286.

7. Ioffe, S. Improved consistent sampling, weighted minhash and l1 sketching. In *Proc. 10th IEEE International Conference on Data Mining 2010* (2010), pp. 246–255.

8. Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In *Proc. 10th European Conference on Machine Learning*, vol. 1398 of *Springer Lecture Notes in Computer Science*. pp. 137–142.

9. Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*, 1 (1964), 1–27.

10. Lazebnik, S., Schmid, C., and Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2006), vol. 2, pp. 2169–2178.

11. Lowe, D. G. Object recognition from local scale-invariant features. In *Proc. 7th IEEE International Conference on Computer Vision* (1999), vol. 2, pp. 1150–1157.

12. Mamani, G. M. H., Fatore, F. M., Nonato, L. G., and Paulovich, F. V. User-driven feature space transformation. *Computer Graphics Forum 32*, 3 (2013), 291–299.

13. Misue, K. Drawing bipartite graphs as anchored maps. In *Proc. Asia-Pacific Symposium on Information Visualisation 2006 (APVis '06)* (2006), pp. 169–177.

14. Misue, K. Anchored map: Graph drawing technique to support network mining. *IEICE Transactions 91-D*, 11 (2008), 2599–2606.

15. Mizuno, K., Wu, H.-Y., and Takahashi, S. Manipulating bilevel feature space for

category-aware image exploration. In *Proc. of the 7th IEEE Pacific Visualization Symposium (PacificVis 2014)* (2014), pp. 217–224.

16. Paulovich, F. V., Eler, D. M., Poco, J., Botha, C. P., Minghim, R., and Nonato, L. G. Piecewise laplacian-based projection for interactive data exploration and organization. *Computer Graphics Forum 30*, 3 (2011), 1091–1100.

17. Perronnin, F. Universal and adapted vocabularies for generic visual categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 7 (2008), 1243–1256.

18. Sato, S., Misue, K., and Tanaka, J. Readable representations for large-scale bipartite graphs. In *Proc. Knowledge-Based Intelligent Information and Engineering Systems*, vol. 5178 of *Springer Lecture Notes in Computer Science*, pp. 831–838.

19. Sivic, J., and Zisserman, A. Video google: a text retrieval approach to object matching in videos. In *Proc. 9th IEEE International Conference on Computer Vision* (2003), pp. 1470–1477.

20. Torgerson, W. S. Multidimensional scaling: I. theory and method. *Psychometrika 17*, 4 (1952), 401–419.

21. Winn, J., Criminisi, A., and Minka, T. Object categorization by learned universal visual dictionary. In *Proc. 10th IEEE International Conference on Computer Vision* (2005), vol. 2, pp. 1800–1807.

22. Yang, J., Yu, K., Gong, Y., and Huang, T. Linear spatial pyramid matching using sparse coding for image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2009).