Semantic Image Clustering Using Object Relation Network

Na Chen and Viktor K. Prasanna

University of Southern California

Abstract. This paper presents a novel method to organize a collection of images into a hierarchy of clusters based on image semantics. Given a group of raw images with no metadata as input, our method describes the semantics of each image with a *bag-of-semantics* model (*i.e.*, a set of meaningful descriptors), which is derived from the image's Object Relation Network [5] - an expressive graph model representing rich semantics for image objects and their relations. We adopt the class hierarchies in a guide ontology as different levels of lenses to view the bag-of-semantics models. Image clusters are automatically extracted by grouping images with the same bag-of-semantics viewed through a certain lens. With a series of coarse-to-fine lenses, images are clustered in a top-down hierarchical manner. In addition, given that users can have different perspectives regarding how images should be clustered, our method allows each user to control the clustering process while browsing, and thus dynamically adjusts the clustering result according to the user's preferences.

1 Introduction

Image clustering is an important tool in processing large collections of images. The goal of image clustering is to organize a large set of images into clusters, such that images within the same cluster have similar meaning. Image clustering provides high-level summarization of large image collections, and thus has many useful applications. For example, clustered web image search results and image repositories are more convenient for users to browse. In addition, the efficiency of image search in large image database can be significantly improved by retrieving clustered image groups rather than individual images.

Many research efforts tackle the complicated problem of image clustering by solving three subproblems. Given a collection of images, first, a set of features are extracted from each image as its description. The features can be low-level visual features (*e.g.*, [17,7,13]), web context features (*e.g.*, [3,11,18]), or region-based features such as the well-known bag-of-words model [15,2,12]. Second, a clustering algorithm (*e.g.*, k-means, NCut, kNN) is applied based on certain distance measurements defined in the feature space, to split the image collection into multiple clusters. Finally, each cluster is labeled with either a text description or a representative image.

Although the previous research succeeds in many applications, we notice two major limitations. First, current visual feature based clustering methods usually

use local features that do not have semantic meanings. Thus, given two images, there is no significant correspondence between their semantic distance and their visual feature distance. These methods risk grouping images with different semantics into the same cluster, which is unsatisfactory from the perspective of human users. Although supervised machine learning approaches can be introduced to reduce the gap between local visual features and image semantics, they may fail when dealing with specific semantics. *E.g.*, they can hardly tell the semantic difference between the ball-playing scenes in Figure 1 left column.

The second limitation of the current image clustering methods is that they usually act as a black box to users, who has no control of the clustering performance. However, we observe that different users can have very different purposes of clustering. *E.g.*, a user focused on ball types wants to group the top two images in Figure 1 together since both of them contain a *soccer ball*, while a user targets at person types wants to group the bottom two images together as they are both about *athletes*. Thus, users should have control of the image clustering process.

We present a novel image clustering method to address the above two issues. Our approach is based on Object Relation Network (ORN) [5], a graph model representing informative and consistent semantics for objects and their relations in an image (*e.g.*, Figure 1 middle column). Given an ORN automatically generated for each image, we propose an image feature model named *bag-of-semantics*, which contains a set of semantic descriptors for the image based on its ORN (*e.g.*, Figure 1 right column). Since the ORN is derived from a *guide ontology*, the class hierarchies in the guide ontology can serve as different levels of lenses to view the bag-of-semantics model. In particular, a lens consists of a set of ontology classes that can be distinguished under it. For example, viewed from a coarse lens involving only *Person* and *Ball* nodes, the bag-of-semantics models for all three images in Figure 1 become the same set $\{Person, Ball\}$. In contrast, viewed through a finer lens involving *Basketball* and *Soccer ball*, the semantic difference in ball types between the bottom image and the other two images can be easily identified. Therefore, we cluster images by grouping images with the same bag-of-semantics viewed through a certain lens. We achieve hierarchical image clustering by going top-down through the class hierarchies in the guide ontology (and thus a series of coarse-to-fine lenses). In addition, user preferences in clustering can be captured by choosing different lenses at certain levels (*e.g.*, splitting *Person* into subclasses and splitting *Ball* into subclasses lead to different clustering results for images in Figure 1). Finally, each image cluster is labeled with its bag-of-semantics under the corresponding lens.

Given that our method is built on some prior work, we state our original contributions as follows:

1. We propose a bag-of-semantics model to describe images for the image clustering problem. The model explicitly reveals the semantics of an image. Thus, our clustering algorithm is guaranteed to group semantically-similar images into the same cluster.

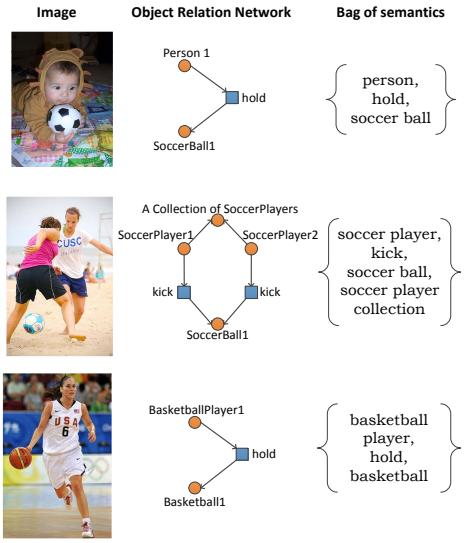

Fig. 1. Images and their bag-of-semantics descriptions automatically generated from our system. ORNs are shown in the middle as intermediate results.

2. We present a top-down hierarchical image clustering algorithm by viewing the bag-of-semantics model through levels of lenses with different semantic granularities, based on the class hierarchies in the guide ontology.
3. We are the first to enable user control in the image clustering problem. We provide a mechanism for users to browse through the image collection and make intuitive adjustment to the clustering results.

2 Related Work

Pioneer image clustering research [17,7,13] extracts low level visual features from input images, and applies different clustering algorithms based on these visual features. These algorithms include distance based clustering [17], Ncut [7], locality preserving clustering [19], and agglomerative clustering [13]. In particular, trees are suggested to be a natural organization of clusters [7]. But in these pioneer efforts, there is no correspondence between the cluster tree and the structure of image semantics.

For web images, textual context is believed to be a useful addition to the visual features. Co-clustering approaches are introduced to integrate visual features and multiple context features such as surrounding text [3,11], links [3,18], and attributes of various data objects [18]. In addition, Jing *et al.* [14] identify semantic clusters related to a given query, and assign the result images to the clusters. These methods work well for specific web applications, but lose generality and accuracy when dealing with images with limited or irrelevant web context.

In computer vision, *image categorization* targets at labeling images with one of a number of predefined categories [6]. Instead of directly using low level visual features (*e.g.*, colors and textures), intermediate representations are frequently introduced to capture image semantics. For example, the well-known bag-of-words model [15,2,12] describes an image as a bag of visual codewords and provides various measurements of image similarity. Another popular intermediate representation consists of image regions created from segmentation. With this representation, the image categorization problem can be formulated as a multiple-instance learning (MIL) problem by viewing an image as a bag of instances [6,1,16].

Although these image categorization methods share some similarities with our approach, we are the first to exploit the relations between objects in images. By exploring the relations between image objects, our bag-of-semantics model can express concrete semantics such as "basketball player" and "soccer player". In addition, we are the first to enable user control in the clustering process.

3 Bag-of-Semantics Model

We model an image as a collection of semantic descriptors for both static image objects and the binary relations between them. In particular, we adopt Object Relation Network (ORN) [5] to capture image semantics. ORN is a graphical model that links objects in an image through meaningful relations (Figure 1 middle column). Guided and constrained by a *guide ontology*, ORN represents the most probable meaning of the objects and their relations, by assigning each graph node to the most probable class in the guide ontology. Therefore, an image can be described by the ontology class assignments in the ORN, *e.g.*, Figure 1 right column. This image description model captures the semantics of both objects and their relations, and thus we call it *bag-of-semantics*.

This hierarchical structure of the semantic descriptors is very useful in our image clustering method, because it can describe an image with semantics from very general level to very specific level. Clustering is achieved by grouping images with the same semantics under certain semantic granularity. In the next section, we formally define *lenses* to control the semantic granularity and present our hierarchical image clustering algorithm based on the bag-of-semantics model.

4 Image Clustering

4.1 Lenses

Lenses characterize the semantic granularity for the bag-of-semantics model. We define a lens as a set of ontology classes $L \subseteq G$, where G is the node set of the guide ontology. Viewed through a lens L, a semantic descriptor s is regarded as its closest ancestor in L, denoted as s_L.

Intuitively, L determines the ontology classes that can be distinguished by the lens. The coarsest lens contains only three general classes, *i.e.*, *Object*, *O-O Relation*, and *Object Collection*. With this lens, every semantic descriptor is

mapped to one of the general classes. Little difference can be found between the descriptors. On the contrary, under a fine lens containing many specific semantic concepts such as *Basketball* and *Soccer Ball*, the corresponding semantic descriptors (*e.g.*, the balls in Figure 1) are expressed with specific concepts and distinguished accordingly.

4.2 Image Clustering with Lenses

Viewed through a lens L, the bag-of-semantics $S(I)$ of an image I are expressed as set $S_L(I) = \{s_L | s \in S(I)\}$. Two images with the same bag-of-semantics expression $S_L(I) = S_L(J)$ are indistinguishable under the lens L. We thus group images with the same bag-of-semantics expression under a certain lens into the same cluster.

At first glance, this clustering algorithm may produce as many as $2^{|L|}$ clusters since $S_L(I) \subseteq L$ and there are $2^{|L|}$ possible subsets of L. However, since ORNs are created following the semantic constraints in the guide ontology, many of L's subsets do not have feasible ORNs, and thus the image clusters corresponding to these subsets are empty.

4.3 Hierarchical Clustering

We propose a top-down hierarchical image clustering algorithm by going through a series of coarse-to-fine lenses. We start with the coarsest lens containing only the general classes and cluster images accordingly. With more specific semantic concepts added to the lens, we divide each cluster into sub clusters according to the refined lens. In particular, we take advantage of the class hierarchies of the guide ontology $\mathbb{G}$. In each lens refinement step, we adopt a *split* operator to a class node f in L that has not been split. The split operator adds f's child class nodes in $\mathbb{G}$ to L, and divides the image clusters according to the refined lens. The hierarchical image clustering algorithm stops when there are sufficient number of clusters.

The pseudo code of the automatic hierarchical image clustering algorithm is shown in Algorithm 1.

4.4 User Control in Image Clustering

Since the bag-of-semantics model carries rich semantics of images, user preferences can be captured by choosing different coarse-to-fine paths for the lens. We design a user control mechanism that allows each user to modify the sequence of class nodes to be *split*. Figure 2 shows an example of a clustering process controlled by a user who is interested in various relations.

In our implementation, the image clustering system first applies the automatic clustering algorithm (Algorithm 1) to generate an initial cluster hierarchy for the user to browse through. In each lens refinement step, the user has the option to participate and choose a class node he thinks to be the most important. Our

Algorithm 1: Hierarchical Image Clustering

Input: Image collection $\mathbb{I}$ with bags-of-semantics $S(\mathbb{I}) = \{S(I)|I \in \mathbb{I}\}$, guide ontology $\mathbb{G}$
Output: Image clusters $\mathbb{C} = \{C_i\}$
Initialization: $L, F \leftarrow$ {visible general classes}
$\mathbb{C} \leftarrow$ clusters of $\mathbb{I}$ generated using L
while $F \neq \varnothing$ *and* $|\mathbb{C}| < \delta$ **do**
  find $f \in F$ with the smallest depth in $\mathbb{G}$
  find f's visible child node set $Child_v(f)$ in $\mathbb{G}$
  $F \leftarrow F \bigcup Child_v(f) \setminus \{f\}$
  if $Child_v(f) \neq \varnothing$ **then**
    $L \leftarrow L \bigcup Child_v(f)$
    foreach $C_i \in \mathbb{C}$ **do**
      divide C_i into sub clusters using L
      replace C_i in $\mathbb{C}$ with its sub clusters

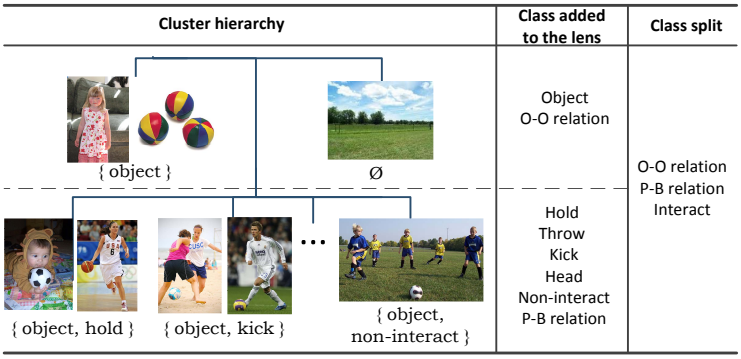

Fig. 2. Different choices of lens result in different cluster hierarchies. Lenses are chosen to cluster images according to relation types and person types in (A) and (B) respectively.

system splits the specified node, finds clusters based on the refined lens, then applies Algorithm 1 to update the subsequent cluster hierarchy that is used for further browsing. Finally, each image cluster is labeled with its bag-of-semantics under the corresponding lens.

5 Experimental Results

The dataset used in our experiment contains over 28,000 images from VOC2011 [9] and ImageNet [8]. We randomly choose 2,000 images for the training process of ORN generation. Our guide ontology contains class hierarchies and constraints for 6 generic object classes (*Person, Bicycle, Motorbike, Horse, Chair, Ball*) and their relation classes. We adopt the detectors in [10] to perform object detection.

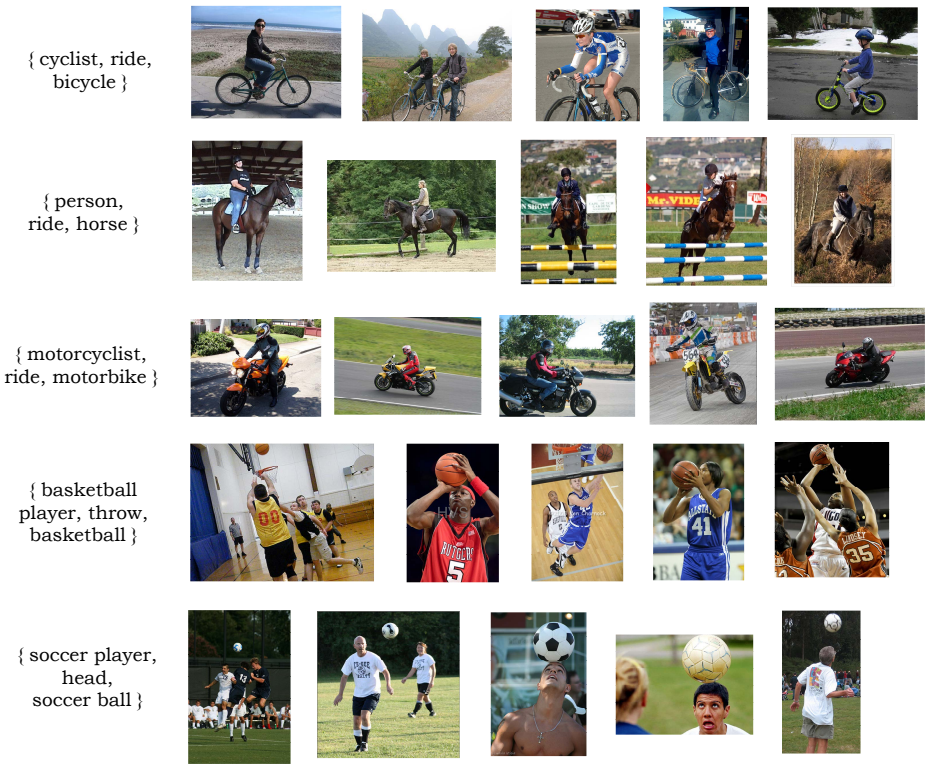

Fig. 3. Several "good" clustering results obtained by using very fine lenses, *i.e.*splitting subclasses of *Person* class and all the relation classes. Each row includes example images and the label of a cluster.

Figure 3 illustrates several "good" results produced by our automatic clustering algorithm, obtained by splitting subclasses of *Person* class and all the relation classes. These result clusters demonstrate that our clustering algorithm has successfully classified semantically-similar images into the same cluster, even though the visual features of some images are quite different from each other.

More results and discussions are available in [4].

6 Conclusion

We presented a hierarchical image clustering method that groups semantically similar images into the same cluster. We proposed a bag-of-semantics model to describe the semantic features of images. Viewed through a series of coarse-to-fine lenses, images with the same bag-of-semantics under a certain lens are clustered in a top-down hierarchical manner. Our method allows each user to control the clustering process while browsing, and dynamically adjusts the clustering result according to his purpose.

References

1. Bi, J., Chen, Y., Wang, J.Z.: A sparse support vector machine approach to region-based image categorization. In: CVPR (2005) 4
2. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via plsa. In: ECCV (2006) 1, 4
3. Cai, D., He, X., Li, Z., Ma, W.Y., Wen, J.R.: Hierarchical clustering of www image search results using visual, textual and link information. In: ACM Multimedia (2004) 1, 3
4. Chen, N., Prasanna, V.K.: A bag-of-semantics model for image clustering. Tech. rep., University of Southern California (August 2012), http://www-scf.usc.edu/~nchen/paper/bos.pdf 7
5. Chen, N., Zhou, Q.Y., Prasanna, V.: Understanding web images by object relation network. In: Proceedings of the 21st international conference on World Wide Web (2012) 1, 2, 4
6. Chen, Y., Wang, J.Z.: Image categorization by learning and reasoning with regions. J. Mach. Learn. Res. (2004) 4
7. Chen, Y., Wang, J.Z., Krovetz, R.: Clue: Cluster-based retrieval of images by unsupervised learning. IEEE Transactions on Image Processing (2003) 1, 3
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. CVPR (2009) 6
9. Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/ index.html 6
10. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE TPAMI 32(9) (2010) 6
11. Gao, B., Liu, T.Y., Qin, T., Zheng, X., Cheng, Q.S., Ma, W.Y.: Web image clustering by consistent utilization of visual features and surrounding texts. In: ACM Multimedia (2005) 1, 3
12. Gemert, J.C., Geusebroek, J.M., Veenman, C.J., Smeulders, A.W.: Kernel codebooks for scene categorization. In: ECCV (2008) 1, 4
13. Gordon, S., Greenspan, H., Goldberger, J.: Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations. In: ICCV (2003) 1, 3
14. Jing, F., Wang, C., Yao, Y., Deng, K., Zhang, L., Ma, W.Y.: Igroup: web image search results clustering. In: ACM Multimedia (2006) 3
15. Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR (2005) 1, 4
16. Liu, Y., Chen, X., Zhang, C., Sprague, A.: Semantic clustering for region-based image retrieval. J. Vis. Comun. Image Represent. (2009) 4
17. Rodden, K., Basalaj, W., Sinclair, D., Wood, K.: Does organisation by similarity assist image browsing? In: Proceedings of the SIGCHI conference on Human factors in computing systems (2001) 1, 3
18. Wang, X.J., Ma, W.Y., Zhang, L., Li, X.: Iteratively clustering web images based on link and attribute reinforcements. In: ACM Multimedia (2005) 1, 3
19. Zheng, X., Cai, D., He, X., Ma, W.Y., Lin, X.: Locality preserving clustering for image database. In: ACM Multimedia (2004) 3