Understanding Human Actions: An Information Network Perspective of View

Bangpeng Yao, Bo Wang, Cherhan Foo

Stanford University

1 Introduction

Understanding human actions is an important yet challenging problem in many research fields, such as psychology [7], linguistics, and computer vision [9]. Comparing with object recognition where researchers have started building hierarchical structures for millions of classes in both natural language processing [5] and computer vision [4], little is known about the nature of human actions (e.g. what makes an action to be "riding a horse"?) as well as the relationship between different actions (intuitively, "playing guitar" should be functionally more similar to "playing violin" than "drinking water"). Interestingly, this follows a similar story as the evolution of human brain and languages. It has been reported that infants learn nouns much earlier than verbs [7]. One hypothesis is that most of the nouns, which corresponds to objects, refer to existing things and entities in the world that are easily accessible to humans. The verbs, which are closely related to actions, on the other hand, tend to describe the concepts that are more abstract and more diffuse in the feature space.

In the field of computer vision, researchers usually regard action understanding as a pure classification problem [9]. Typically, people have a list of pre-defined actions, and then collect a set of images and annotate the humans that are doing one of the actions. Then the only goal is to design a classifier which is able to automatically assign a class label to each human image.

Fig. 1. Human actions lie in a continuous space, where the human pose can change continuously and convey different meanings.

Fig. 2. Human actions are highly related to human poses and objects. In the left two images, "riding bike" and "riding horse" are functionally related and the humans have similar poses. In the right two images, both humans are "playing tennis". The poses of the two humans are very different, but they both interact with the same object (tennis racket).

Fig. 3. The human actions can be different even when the humans interact with the same object. We observe that human poses in "riding bike" are usually similar, whereas the poses in "fixing a bike" are very different.

However, the tasks of understanding human actions is more complicated than simply performing classification. As shown in Figure 1, human actions lie in a continuous space where one action can be more closer to a second action than the others. Furthermore, there are relationships between different actions. As shown in Figure 2, the human poses and objects that the humans are interacting with can make one action related to the other. Furthermore, even within the same action, there are some special distributions of human poses as well as the spatial relationships between objects and humans, and those distributions are different even when the humans are interacting with the same object, as shown in Figure 3.

In this project, we study the properties of different human actions, as well as the relationship between different action images, instead of treating action recognition as a pure classification problem with human-provided class labels. We are interested in the following three questions:

- How to model an action image and measure the similarity between two images?

- How to measure the similarity between two action classes, hence discover a hierarchical relationship between all actions?
- Without class labels, how can we cluster action images in an unsupervised way?

Specifically, we use an informative network analysis approaches to achieve the above goals. We form an "action network" where each node is an action image[1]. There is an edge between two images if they are semantically related to each other, e.g. where the human poses are similar, or the humans interacting with the same object in similar ways, etc.

2 Prior Work

In this work, we consider all the action images in a network, where each node is an action image. Representing action images in a network allows us to jointly consider the relationships of all the images. The distance between two nodes in the network reflects the degree of relevance of the two actions, e.g. two action images are highly related if they are directly connected in the network. Under this distance measure, the distance between two action classes can be measured by the average distance between pairs of actions of this two classes. Furthermore, we can also use network clique detection algorithms, such as [3, 6], to automatically detect the "communities" in the network.

Each node in the network, an action image, is represented by a set of units of human poses and objects that are interacting with the humans. As shown in Figure 2 and Figure 3, the objects and human poses together define the complex relationships between different action images. Representing human actions as objects and human poses has been studied in [9]. However, in that paper, the goal is to simply classify human actions into a pre-defined list of action classes, without considering the relationships between different action classes.

3 Algorithm

3.1 Action Images Represented by Poselets and Objects

As shown in Figure 1 and Figure 3, the relationships between different action images can be measured by the human poses and objects. Intuitively, two actions are similar if they share the same set of objects with similar spatial configurations between humans, or correspond to similar human poses. We therefore decompose each action image into a set of poselets [2] and objects.

The poselets are a set of local human poses that are tightly consistent in terms of layout of human body parts, as shown in Figure 4. Each poselet can answer a pose-related question. For example, "is the human sitting?", "Is the human's hands crossing?", etc. Given a set of images with annotations of keypoints (as shown in Figure 5) of the human bodies, we use a clustering method to obtain

[1] We assume that there is only one human in each action image.

Fig. 4. Examples of poselets. Images in each row correspond to one poselet.

250 poselets. The clustering method we use is very similar to that in [2]. We repeatedly draw random rectangles on the action images with different sizes and locations, where each rectangle is a candidate poselet. For each rectangle, we search all the other images to find the rectangles where the configurations of keypoints are similar to the reference rectangle. If this search leads to a similar rectangle in an image, we say that this image contains this poselet. From the 2000 candidate poselets, we merge the ones that are similar to each other, and delete the poselets which commonly happen on almost all the images (e.g. an upright torso). We also merge the poselets which happen in a very small number of images. After all the processing, we obtain 250 poselets.

For the objects, we specify a list of objects that interact with the humans (such as horse, bike, pen, chair, etc), and draw a tight bounding box on each object if the object appears in the image. Therefore we have the information of whether a specific object appears in each image. Furthermore, since humans might interact with the same object in many different ways (such as “riding a bike” versus “fixing a bike”, “riding a horse” versus “feeding a horse”), we also consider different spatial relationships between humans and objects. Based on the locations of bounding boxes of humans and objects, we consider five different object locations with respect to humans: overlapping, top, bottom, left, and right.

Therefore, each image can be represented as a 500-dimensional binary vector, where 250 elements describe whether a specific poselet is contained in the image, while other 250 elements describe whether an object appears in the image and what is the spatial relationship between this object and the human.

Fig. 5. Demonstration of our annotation. The magenta bounding box indicates the object, while the red dots indicate the key points in the human body. The key points we consider are: top and bottom points of head and torso; top, bottom points and elbow of left and right arms; top, bottom, and knee of legs.

Fig. 6. The left-most figure illustrates our division of the image locations with respect to the human. The right two images show the location of the same object in different actions.

3.2 The Stanford-40 Dataset and Action Network

We perform experiments on the Stanford-40 Action Dataset [9], where there are 40 actions with 180-300 images per action. We use half of the total images on all the classes for our experiment. The keypoints of human body parts and object bounding boxes are annotated as in Figure 5. Based on the annotations, we obtain a 500-dimensional binary vector for each image, as described in Section 3.1. Each element of the vector corresponds to a poselet or an object with a specific relationship to the human.

We then form the action network in the following way. Each image is a node in the network. There is an edge between two nodes if they share more than three poselets or one object (with the same human-object spatial relationship), or both. This gives us a network that has 4766 nodes and 227099 edges. Note that here we use an undirected unweighted network to make the computation more tractable. We can also form weighted action networks by adding larger weights to the pairs of images which are very similar to each other.

3.3 Action Similarity and "Action Community" Detection

Based on the action network defined above, here we describe our approach for measuring the similarity between two action classes and how to automatically detect "action communities" from the data.

Our method for measuring the similarity between two actions is straightforward. For every pairs of images, we compare the shortest path between the two images in the network. Then the distance between two action classes is the average distance between all possible pairs of images, one from each class.

Measuring the similarity between different classes provides us more information than just the class labels. We can know which two actions are more related than the others. However, as we have discussed in Section 1, actions refer to abstract concepts, and there is not an abstract way to define the partition of the action space. Therefore we also want to analyze how to automatically partition the action images into a set of classes, without annotations of action classes. Our goal is to find a number of clusters of images, where each cluster consists of a number of images within which the node-node connections are dense, and the edges to nodes in other communities are less dense. This is very similar to the community detection problem in network analysis. We use the Girvan-Newman algorithm [8] to solve this problem. The Girvan-Newman algorithm repeats the following two steps until no edges are left.

- Calculate the betweenness of all pairs of edges;
- Remove edges with highest betweenness.

The Girvan-Newman algorithm gives us a way of clustering action images on the action network. But how to determine the number of clusters is an open problem. We empirically set the number of clusters to be in the range of $50 \sim 200$. The clustering procedure terminates once any number of clusters in the desired range is obtained.

4 Results and Findings

Figure 7 demonstrates the confusion matrix of the distance between each pair of action classes. We set the distance between the same action class to be 0 (the diagonal of the matrix). From the matrix, we have the following observations.

- The distance between different pairs of action classes vary a lot. There are some actions that have small distance, e.g. "writing on a book" and "reading book" (the same set of object "book"), "blowing bubbles" and "taking photos" (similar human poses), etc. There are also actions that are completely irrelevant, e.g. "climbing wall" and "rowing a boat".
- The actions of "rowing a boat" and "walking a dog" have large distance with almost all the other actions. The reason might be that "rowing a boat" does not share any object or human pose with other classes, and there are no distinct human poses in the action of "walking a dog".

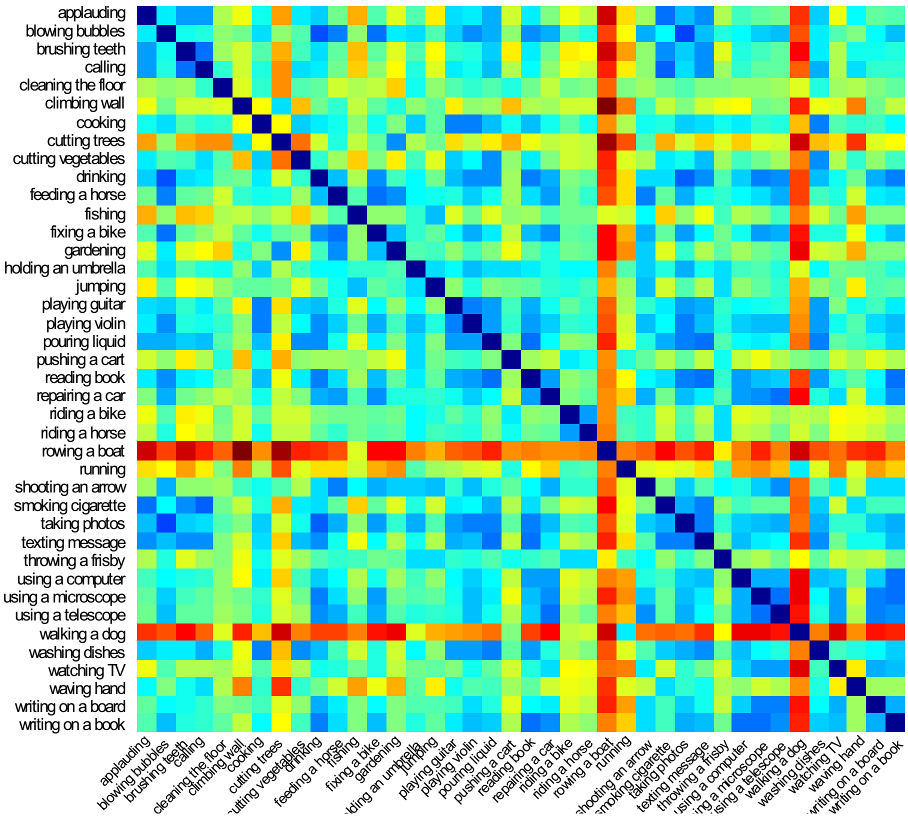

Fig. 7. The distance between pairs of actions. Blue color indicates small distance, while red color indicates large distance. We set the distance between the same action class to be 0.

– It is not necessary for two actions where humans interact with the same object have small distance. We observe that the actions "riding a horse" and "feeding a horse", "riding a bike" and "fixing a bike", do not have very small distance. The reason is although those actions contain the same object, the spatial relationship between the human and the object is different. However, for the actions that are functionally related, e.g. "using a microscope" and "using a telescope", "playing guitar" and "playing violin", are very related, because of the similar human poses.

Figure 8 illustrates our action image clustering results. We visualize the number of images that belong to each cluster for each action class. We observe that most of the images belong to some specific action clusters. For some clusters, the number of images is very small (e.g. the first several rows in the figure). On the one hand, this figure shows that clustering action images based on objects and human poses only is not a good choice, because we cannot obtain the desired

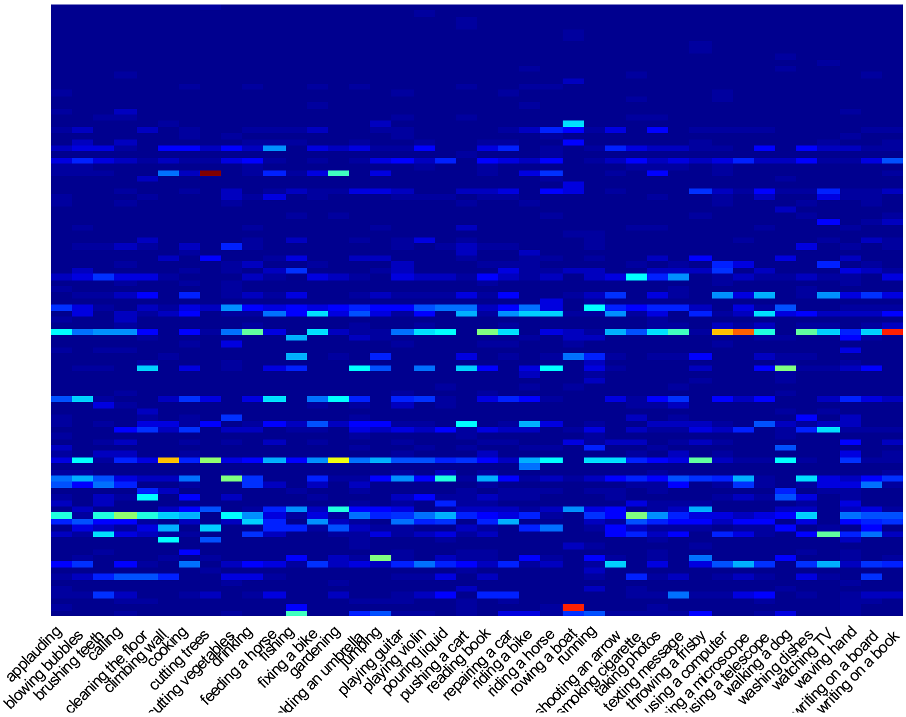

Fig. 8. The action clustering results. Each row is a cluster, each column is the number of images of a specific action in different classes. Red color indicates large number, while blue color indicates small number.

clustering structure from Figure 8. On the other hand, however, there are some correlations between Figure 8 and Figure 7. We observe that most of the images of "rowing a boat" belong to one cluster, while Figure 7 shows that "rowing a boat" has large distance to all the other actions.

References

1. American Time Use Survey (ATUS) Activity Lexicon. Bureau of Labor Statistics, 2010.
2. Bourdev, L., Malik, J.: Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. International Conference on Computer Vision (ICCV), 2009.
3. Clauset, A., Newman, M., Moore, C.: Finding Community Structure in Very Large Networks. Physics Review E, volume 70, 2004.
4. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.
5. Fellbaum, C.: An Electronic Lexical Database. Bradford Books, 1998.
6. Flake, G., Tsioutsiouliklis, K., Tarjan, R.: Graph Clustering Techniques based on Minimum Cut Trees. Technical Report 06, NEC, Princeton, NJ, 2002.

7. Gentner, D.: Why Nouns Are Learned Before Verbs: Linguistic Relativity Versus Natural Partitioning. Language Development, Volume 2: Language Thought and Cluture. Hillsdale, NJ, 1982.
8. Girvan, M., Newman, M.: Community Structure in Social and Biological Networks. Proceedings of the National Academy of Sciences, 99:7821-7826, 2002.
9. Yao, B., Jiang, X., Khosla, A., Lin, A., Guibas, L., Fei-Fei, L.: Human Action Recognition by Learining Bases of Action Attributes and Parts. Internatioinal Conference on Computer Vision (ICCV), 2011.