Understanding action recognition in still images

Deeptha Girish
University of Cincinnati
Cincinnati, OH 45220, USA
girishde@mail.uc.edu

Vineeta Singh
University of Cincinnati
Cincinnati, OH 45220, USA
singhvi@mail.uc.edu

Anca Ralescu
University of Cincinnati
Cincinnati, OH 45220, USA
anca.ralescu@uc.edu

Abstract

Action recognition in still images is closely related to various other computer vision tasks such as pose estimation, object recognition, image retrieval, video action recognition and frame tagging in videos. This problem is focused on recognizing a person's action or behavior using a single frame. Unlike action recognition in videos - a relatively very well established area of research where spatio-temporal features are used, these are not available for still images, making the problem more challenging. In the present work only actions that involve objects are considered. A complex action is broken down into components based on semantics. The importance of each of these components in action recognition is systematically studied.

1. Introduction

Video based action recognition is relatively a well-established and well-studied area of research, whereas still image based recognition is comparatively less studied. It has gained a lot of attention in the past few years with increasing number of images available from social networks. Since motion cannot be easily estimated from a still image and spatio-temporal features cannot be used for characterizing the action, action recognition in still image remains a challenging problem. Though it is more intuitive and easier to determine action in videos, it is possible and very useful to recognize actions in static images. Many action categories can be depicted unambiguously in single images (without motion or video signal), and these actions can be understood well based on human perception. For such action categories a single frame is sufficient to accurately classify actions. This evidence supports the development of computational algorithms for automated action analysis and recognition in still images.

Still image based action recognition has many useful applications. It can be used for surveillance, robotic applications, human computer interaction applications, annotating images using verbs, searching an image database using verbs, searching images online based on action queries, frame tagging, searching in videos and understanding the functionality of an object, and video frame reduction in video based activity recognition. Long video sequences can be reduced to fewer frames for action representation, thus decreasing redundant information without compromising on the accuracy.

The main challenges for action recognition in still images are loss of spatio-temporal features, background clutter, high intra-class variance and low inter-class variance among some action classes, change in background lighting and variation in person pose. Spatio-temporal features is the most important feature used to characterize actions in videos. In case of images, the temporal information is lost which makes it significantly harder to represent an action.

Action Recognition in still images is related to other important computer vision tasks like object recognition, video based action recognition, pose estimation, scene recognition and image retrieval. For some tasks such as image retrieval and video based action recognition, still image based action recognition is the preliminary step. The results from still image based action recognition are used as a feature and combined with other features extracted depending on the problem statement. On the other hand, object recognition is used as a preliminary step for action recognition. Off the shelf object detectors are often used to get the class labels of all the objects present in the image, their co-occurrence is modeled and used as a feature to perform still image based action recognition. Other computer vision tasks such as pose estimation and scene understanding are also closely related to still image action recognition. Many papers have used pose estimation of the person in the image and the scene in which the action is being performed as an input to action recognition. The action recognition model is trained with pose and scene features to a certain degree of accuracy. The trained still image action recognition model is in turn used as an input to achieve a more accurate pose estimation and scene understanding model. This forms a loop where one task is used to improve the performance of the other.

In this work only actions that involve a person manipu-

lating an object are considered. The aim is to break down an action into smaller semantic components and understand the importance of each of these components in action recognition.

1.1. Related Work

Human body, body parts, action-related objects, human object interaction, and the whole scene or context are the most popular high level cues used in human action recognition in still images. These cues characterize different aspects of human actions [13]. Wang et al. [18] exploited the overall coarse shape of human body in the image. The shape was represented as a collection of edge points obtained via canny edge detector [3]. The shape is used as features to cluster and label images into different actions. Body pose is also an important cue for action recognition. Ikizler et al. [9] used the body poses extracted from images using edge and region features used to construct deformable models using the Conditional Random Field (CRF). Yao et al. [20] used a variation of random forests to search the useful, discriminative patches from the human body region for action recognition. Critical patch information is also represented in the form of a saliency map [17].

Poselets [2] extracted from body parts, capture salient body poses specific to certain actions. Body parts for have been analyzed using poselets for still image based action recognition in Maji et al [12] and Zheng et al. [22]. Raja et al. [16] considered a graphical model containing six nodes encoding positions of five body parts and the action label.

A lot of actions performed by humans involve objects, thus is is useful to consider relevant and related objects for action characterization. Prest et al.[15] used the notion of objectness to calculate the probability of a certain patch being an object. Objectness predicts if image patches belong to an object irrespective of the object class.

Yao et al. [19] used a part based model composed of objects and human poses. The related objects are either manipulated by persons (e.g., a bike is ridden by a person) or related to the scene context of the action (e.g., the grass in the scene of "horse riding in grassland"). The attributes are linguistically related descriptions of human actions. The parts are composed of objects and human poses. Attributes and parts are used as action bases to model actions in still images [8]. In [10], Le et al. input images are broken down into recognized objects and a language model is used to enumerate all possible actions when the objects are used in different configurations. Some methods model co-occurence of objects to characterize actions, while others have integrated scene information with the object information for action recognition.

The configuration of human and objects is very specific to each action. Apart from co-occurrence of objects, the interaction of humans and objects can be modeled separately. Features like relative size, relative angle and relative distance can be used to characterize human-object interaction. Desai et al. [5] used contextual information for action recognition, such as object layout obtained by their discriminative models [6]. Maji et al. [12] learned a mixture model of the relative spatial locations between the person bounding boxes and the object bounding boxes in still images. For each object type, they fit a two component mixture model of the predicted bounding box to model various relative locations between the person and the object [8].

Shape Context proposed by Belongie and Malik [1] for extracting and matching shape feature used for segmenting and matching the human and object contours. GIST or spatial envelope proposed by Oliva and Torralba [14] a holistic representation of the scene that captures spatial properties, is used to represent the scene along with other features to aid action recognition.

Generative models learn the distribution of various actions belonging to different classes. Li and Fei Fei [11] used a generative model based on a hierarchical structure for action recognition using spatial and appearance features. Gkioxari et al. [7] used a fast-CNN based model for modeling the human object interaction. They represent every image as a triplet of human verb and object. They hypothesize that the appearance of a person like their pose and action is helpful in localizing the object they are interacting with. Their model learns to jointly predict human, the action performed by the human and the object the human is interacting with in an end to end system called the interact net. Another paper by Yu Zhang et al. [21] argues that human and object bounding boxes are not required to detect actions in still images. They propose an approach that uses VGG net to extract CNN features and a Gaussian mixture model to detect actions in images with minimum annotation efforts. They divide this into two parts: to use selective search to find object proposals and object parts and also find a detailed shape of human- object interaction parts. The second goal is to use these interaction features to make a prediction on the activity.

1.2. Experiment

Since the current study is focused only on actions that involve objects, a custom dataset with eight classes is created. Images for this dataset were chosen from different sources. Some images for a couple of action classes were directly chosen from existing datasets such as Stanford 40 Action dataset [19] and Willow dataset [4] if they had one person performing an action using an object. Other images were scrapped from google search engine. The custom dataset has two hundred images for eight action categories including: *drinking*, *fixing*, *phoning*, *pouring*, *reading*, *riding*, *sitting* and *sleeping*.

In this experiment YOLO version 2 trained on MS

Figure 1. Examples from custom dataset

.

COCO dataset that has 1000 object categories is used to find regions of interest in the image. An action is broken down into fount parts; human region, object region, interaction region and union region. YOLO is used to detect humans and objects in each image in the dataset. The custom dataset is carefully constructed to ensure only object categories in MS COCO dataset appear in the images. YOLO detection gives the bounding box and the class labels. If multiple objects are detected in an image, the object-person combination with highest IOU (intersection over union) is considered.

Only the bounding box coordinates are used in this experiment. Class label information is only used to record if the object under consideration is a human or not. The reason is to understand *what kind of object* is used for a particular action. We are not interested in the object itself. The aim is not to encode co-occurrence of the objects in the feature vector representing the action image, rather it is to generalize among different object categories, encode the common properties among them and evaluate their importance in action classification. Using object category information will focus on co-occurrence rather than specific object properties. For example, we do not want to encode that if there is a couch and a person or a chair and a person in the image the action being performed is sitting. Instead, we would like to encode that an there is an object with a flat surface and a person interacting in a specific way, therefore the action being performed is sitting.

Person and object bounding box coordinates are used to find the intersection region and the union region. Intersection region focuses on the parts of the object and person involved in the action and it is the core of the action. Same object with a different interaction leads to a different action. Intersection region represents the interaction between the person and the object involved in the action. Therefore, characterizing the interaction is essential for action recognition. Union region focuses on the entire action region at once without considering the background.

Interaction region is the overlap between the object and person bounding boxes and union bounding box is the joint area of the object and person bounding boxes. If person bounding box P has coordinates $xp1$, $yp1$, $xp2$ and $yp2$ and object bounding box O has coordinates $xo1$, $yo1$, $xo2$ and $yo2$, the interaction region and the union regions are obtained using the following equations.

Interaction bounding box I is defined by $xi1$, $yi1$, $xi2$ and $yi2$, where

$$xi1 = \max(xp1, xo1)$$
$$yi1 = \max(yp1, yo1)$$
$$xi2 = \min(xp2, xo2)$$
$$yi2 = \min(yp2, yo2)$$

Union bounding box U is defined by $xu1$, $yu1$, $xu2$ and $yu2$, where

$$xu1 = \min(xp1, xo1)$$
$$yu1 = \min(yp1, yo1)$$
$$xu2 = \max(xp2, xo2)$$
$$yu2 = \max(yp2, yo2)$$

Using the above equations and the bounding box coordinates generated by YOLO for object and person, the interaction bounding box and the union bounding box is obtained. CNN codes (Activations from FC7 layer) obtained from AlexNet trained on ImageNet are extracted from each of these regions. Each of these components are presented using a 4096 dimensional vector, the high level features obtained from a pre-trained CNN. These 4096-dimensional vectors are concatenated to represent one image in the dataset. A classifier is trained on these concatenated features to perform action classification. The algorithm for action recognition in still images is as follows.

1.2.1 Custom interaction feature

To specifically encode spatial relationships between pairs of action components a custom interaction feature is introduced. The interaction between any two action components $c1$ and $c2$ is defined as follows.

Interaction (c1, c2) = [area (c1), area(c2), $\frac{area(c1)}{area(c2)}$*, distance(c1, c2), angle feature(c1, c2)]*

where $distance(c1, c2)$ is the euclidean distance between the bounding box centers of components $c1$ and $c2$.The enter of the bounding box of a component c with bounding box coordinates $x1$, $y1$, $x2$ and $y2$ is defined as

$$Center(c) = \left[\frac{x1 + x2}{2}, \frac{y1 + y2}{2}\right]$$

Distance between components $c1$ and $c2$ with centers $(cx1, cy1)$ and $(cx2, cy2)$ is defined as follows.

$$distance(c1, c2) = \frac{\sqrt{(cx1 - cx2)^2 + (cy1 - cy2)^2}}{\sqrt{(xu2 - xu1)^2 + (yu2 - yu1)^2}}$$

The angle between two components $c1$ and $c2$ with centers $(cx1, cy1)$ and $(cx2, cy2)$ is defined as follows.

$$angle(c1, c2) = \arctan \frac{cy2 - cy1}{cx2 - cx1}$$

The angle feature is an eight dimensional sparse vector. Angles from 0 to 360 is divided into an eight equal bins. The bin the $angle(c1, c2)$ falls into is represented by 1 and the rest are represented by 0. for example if the angle between two components is 20 degrees, the angle vector will be [1, 0, 0, 0, 0, 0, 0, 0] and if the angle is 170 degrees, the angle vector will be [0, 0, 0, 1, 0, 0, 0, 0]. The angle vector makes the feature vector more robust to slightly different poses because different people may perform the same action differently though the overall pose remains the same. Also, calculating the angle between the component centers makes the feature vector rotation invariant.

The distance between centers of any two regions is normalized by the diagonal of the union region. The union region represents the area of the image in which the action is taking place, therefore normalizing the distance by the union diagonal makes the feature invariant to scale.

The 36-dimensional custom interaction feature is defined as follows.

custom feature = [interaction(person,object), interaction (person, object-person intersection region), interaction (object, object-person intersection region)]

Figure 2 illustrates the workflow of this method.

The combination of feature components used in this experiment is listed below.

- **FC7-all components** This is a combination of CNN codes (FC7 activations) of each of the components. [person, object, interaction, union]
- **FC7-all components + custom relative feature** This is a combination of CNN codes (FC7 activations) of each of teh components and the custom interaction/relative feature. [person, object, interaction, union, custom relative feature]

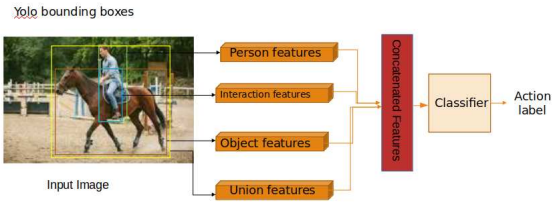

Figure 2. Action recognition algorithm flowchart

.

- **Custom relative feature** Only the 36-dimensional custom relative feature is used in this case.
- **FC7-object** CNN codes of object region.
- **FC7-person** CNN codes of person region.
- **FC7-interaction** CNN codes of interaction region.
- **FC7-union** CNN codes of union region.
- **FC7-object + custom relative feature** CNN codes of object region concatenated with custom relative feature.
- **FC7-person + custom relative feature** CNN codes of person region concatenated with custom relative feature.
- **FC7-object + FC-interaction** CNN codes of object region concatenated with CNN codes of interaction.
- **FC7-person + FC-interaction** CNN codes of person region concatenated with CNN codes of interaction.
- **conv5-object** activations extracted from the conv5 layer of Alexnet for object regions

2. Results

The results evaluation of each combination of component features for action classification are presented below. SVM and logistic regression has been used to perform classification. Accuracy of classification using both these classifiers is reported.

Figure 3 tabulates the classification accuracies of when different component feature combinations to represent action images for classifiers SVM and logistic regression.

Figure 4is a visual representation of the classification accuracies obtained using feature combinations for better comprehension and comparison.

It can be seen from the results that the combination FC7-all components + custom relative feature with logistic regression gives the best classification accuracy of 81.97%. This is not surprising because all components are used along with the custom feature which provides a richer description of the image. This result shows the approach of breaking

Feature combinations	SVM	Logistic Regression
FC7-all components	79.06	80.81
FC7-all components + custom relative feature	79.66	**81.97**
Custom relative feature	53.48	48.25
FC7-object	63.95	64.53
FC7-person	64.32	65.82
FC7-interaction	62.20	62.25
FC7-union	69.76	73.25
FC7-object + custom relative feature	64.21	65.69
FC7-person + custom relative feature	68.02	66.27
FC7-object + FC7-interaction	64.53	68.02
FC7-person + FC7-interaction	76.74	73.25
Conv5-object	65.11	61.62

Figure 3. Classification accuracy (in percentage) for Action Recognition for all component feature combinations

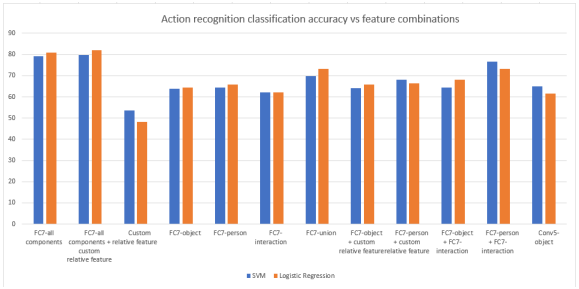

Figure 4. Comparison of classification accuracy for Action Recognition for all component feature combinations

.

down the image into individual components, extracting features from them and then processing them together is a good characterization of an action image.

In general, logistic regression performs marginally better than SVM. FC7-all components gives the second best accuracy (80.81%) which is just slightly lower than FC7-all components + custom relative feature. This shows that when all components are used, the custom relation feature that encodes relative spatial relationships between the components of action does not add much useful information. The custom relation feature by itself produces least classification accuracy indicating that just encoding relative spatial relationships is not enough to classify actions.

FC7-union gives the third best classification accuracy (73.25%). This is very interesting because the FC7-union and FC7-all components cover the same area in the the image, but FC7-all components results in much better accuracy. This proves the hypothesis that separately featurizing action components gives a better and more complete representation that looking at the action as a whole. Separating the action components and then processing them together increases the focus on each of the components while capturing their inter-dependencies and spatial relationships. This makes the model more robust to variations in the dataset leading to better generalization.

Using more features has led to better classification, which is expected. Among the individual regions the union region gives the best results because it has the largest field of view that captures the entire action. The person region does next best followed by object and interaction. It is interesting that the person features perform better than the object features. One reason could be that most objects are partially occluded. Therefore, the entire object is not used for extracting features and depending on the kind of occlusion, forming a general template for objects is harder. Another reason could be that one action uses many different object categories. Objects of the same class also show a lot of variation. Intra-class variance in objects is much higher as compared to person pose for a particular action. Number of ways a person can perform an action is much lower than the variety of objects that can be used for performing an action. Therefore, person is a better cue for action recognition than objects.

Among combination of two regions, FC7-person + FC-interaction gives the best results followed by FC7-person + custom relative feature, FC7-person + FC-interaction and FC7-person + custom relative feature. This shows that person is a better cue than object and using CNN codes for representing interaction is better than the custom relation feature. The difference in accuracies between them is not much considering the difference in the feature dimension between CNN codes for interaction region (4096 dimensions) and custom relation feature (36 dimensions).

Another interesting observation is that using CONV5 activations for objects produces better results than using FC7 activations for objects. This shows that CONV5 features are able to generalize objects better. They are more robust to intra-class variations and better capture the 'kind' of object that is needed to perform a particular action than FC7.

The results evaluation of importance of each combination of component features for recognizing actions based on clustering is discussed in this section. K-means with 8 clusters is used in this study. Normalized mutual information (NMI) is used to evaluation the clusters formed.

Table 1 tabulates the NMI scores for k-means clustering for all component feature combinations. The trends and patterns are similar to classification results. These results bolster our hypothesis that action is made up of individual components and treating them separately and then combining them makes the model more robust.

Figure 5 is a visual representation of the classification

Feature combinations	k-means
FC7-all components	**0.3944**
FC7-all components + custom relative feature	0.2911
Custom relative feature	0.1377
FC7-object	0.2329
FC7-person	0.2049
FC7-interaction	0.1807
FC7-union	0.3015
FC7-object + custom relative feature	0.2438
FC7-person + custom relative feature	0.2202
FC7-object + FC7-interaction	0.2466
FC7-person + FC7-interaction	0.2408
Conv5-object	0.1945

Table 1. Clustering evaluation based on Normalized Mutual Information (NMI) score for all component feature combinations

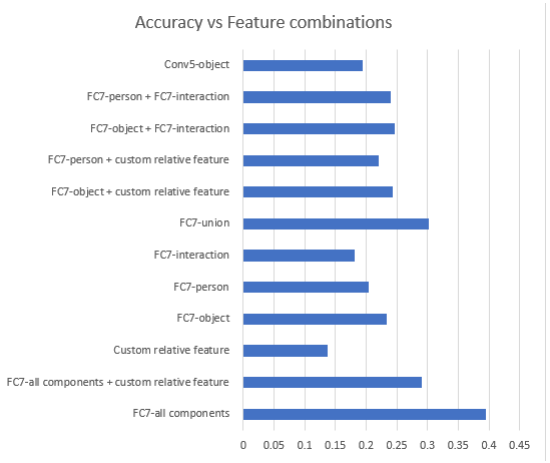

Figure 5. Comparison of accuracy for Action Recognition for all component feature combinations

.

accuracies obtained using feature combinations for better comprehension and comparison.

However, some results contradict the results obtained from the classification based experiment. FC7-all components + custom relative feature performed worse than FC7-all components. CONV5-object performed worse than FC7-object. In both the cases, the worse performing feature has considerably larger dimensions as compared to the better performing feature. Euclidean distance in higher dimensions are less efficient and k-means works well for circular clusters. This artifact of k-means with euclidean distance might be the reason for the disagreement in the classification and clustering result. Another reason could be that the custom relation feature is a sparse vector which might negatively impact the calculation of euclidean distance which is the basis of k-means algorithm.

It is also interesting to note that object component perform slightly better than person component. One reason could be the way NMI score is calculated. NMI has a preference for solutions with more clusters, therefore more variance in objects rather than person might be resulting in higher NMI scores. Another reason for this could be that the way the object features distribute themselves in the object feature space and the distribution of person feature in person feature space is different. Since appearance of objects across images of different objects is much different than the appearance of a person across classes, clustering objects might be easier. Clustering finds natural groups but a classifier can be trained to discriminate groups of data.

3. Conclusion

In this paper a complex action in a still image is broken down into semantic constituents and the importance of each constituent for action recognition is evaluated. Feature importance is evaluated using classification and clustering techniques.

References

[1] Serge Belongie, Greg Mori, and Jitendra Malik. Matching with shape contexts. In *Statistics and Analysis of Shapes*, pages 81–105. Springer, 2006.

[2] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1365–1372. IEEE, 2009.

[3] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[4] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. 2010. updated version, available at http://www.di.ens.fr/willow/research/stillactions/.

[5] Chaitanya Desai, Deva Ramanan, and Charless Fowlkes. Discriminative models for static human-object interactions. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 9–16. IEEE, 2010.

[6] Chaitanya Desai, Deva Ramanan, and Charless C Fowlkes. Discriminative models for multi-class object layout. *International journal of computer vision*, 95(1):1–12, 2011.

[7] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018.

[8] Guodong Guo and Alice Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361, 2014.

[9] Nazli Ikizler, R Gokberk Cinbis, Selen Pehlivan, and Pinar Duygulu. Recognizing actions from still images. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[10] Dieu Thu Le, Raffaella Bernardi, and Jasper Uijlings. Exploiting language models to recognize unseen actions. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 231–238, 2013.

[11] Li-Jia Li and Li Fei-Fei. What, where and who? classifying events by scene and object recognition. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.

[12] Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR 2011*, pages 3177–3184. IEEE, 2011.

[13] Koji Miyajima and Anca Ralescu. Spatial organization in 2d segmented images: representation and recognition of primitive spatial relations. *Fuzzy Sets and Systems*, 65(2-3):225–236, 1994.

[14] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[15] Alessandro Prest, Cordelia Schmid, and Vittorio Ferrari. Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):601–614, 2011.

[16] Kumar Raja, Ivan Laptev, Patrick Pérez, and Lionel Oisel. Joint pose estimation and action recognition in image graphs. In *2011 18th IEEE International Conference on Image Processing*, pages 25–28. IEEE, 2011.

[17] Gaurav Sharma, Frédéric Jurie, and Cordelia Schmid. Discriminative spatial saliency for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3506–3513. IEEE, 2012.

[18] Yang Wang, Hao Jiang, Mark S Drew, Ze-Nian Li, and Greg Mori. Unsupervised discovery of action classes. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1654–1661. IEEE, 2006.

[19] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *2011 International Conference on Computer Vision*, pages 1331–1338. IEEE, 2011.

[20] Bangpeng Yao, Aditya Khosla, and Li Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR 2011*, pages 1577–1584. IEEE, 2011.

[21] Yu Zhang, Li Cheng, Jianxin Wu, Jianfei Cai, Minh N Do, and Jiangbo Lu. Action recognition in still images with minimum annotation efforts. *IEEE Transactions on Image Processing*, 25(11):5479–5490, 2016.

[22] Yin Zheng, Yu-Jin Zhang, Xue Li, and Bao-Di Liu. Action recognition in still images using a combination of human pose and context information. In *2012 19th IEEE International Conference on Image Processing*, pages 785–788. IEEE, 2012.