On Recognizing Actions in Still Images via Multiple Features

Fadime Sener[1], Cagdas Bas[2], and Nazli Ikizler-Cinbis[2]

[1] Computer Engineering Department, Bilkent University, Ankara, Turkey
[2] Computer Engineering Department, Hacettepe University, Ankara, Turkey

Abstract. We propose a multi-cue based approach for recognizing human actions in still images, where relevant object regions are discovered and utilized in a weakly supervised manner. Our approach does not require any explicitly trained object detector or part/attribute annotation. Instead, a multiple instance learning approach is used over sets of object hypotheses in order to represent objects relevant to the actions. We test our method on the extensive Stanford 40 Actions dataset [1] and achieve significant performance gain compared to the state-of-the-art. Our results show that using multiple object hypotheses within multiple instance learning is effective for human action recognition in still images and such an object representation is suitable for using in conjunction with other visual features.

1 Introduction

Recognizing actions in still images has recently gained attention in the vision community due to its large applicability to various domains. In news photographs, for example, it is especially important to understand what the people are doing from a retrieval point of view.

As opposed to motion and appearance in videos, still images convey the action information via the pose of the person and the surrounding object/scene context. Objects are especially important cues for identifying the type of the action. Previous studies verify this observation [2–4] and show that identification of objects play an important role in action recognition.

In this paper, we approach the problem of identifying related objects from a weakly supervised point of view and explore the effect of using Multiple Instance Learning(MIL) for finding the candidate object regions and their corresponding effect in recognition. Our approach does not use any explicit object detector, or part/attribute annotation during training. Instead, multiple object hypotheses are generated via objectness measure [5]. We then utilize a MIL classifier for learning the related object(s) amongst the noisy set of object region candidates.

Besides the features extracted from candidate object regions, we evaluate various features that can be utilized for effective recognition of actions in still images. In our evaluation, we consider facial features in addition to features extracted within person regions and also features that describe the global image

characteristics. We evaluate how much each proposed representation contribute to the recognition of particular actions.

We test our approach on the Stanford 40 actions dataset [1]. Our results show that the MIL framework over the candidate object hypotheses is quite successful and achieves better recognition performance compared to the state-of-the-art part and attributes based model of [1].

The remaining of the paper is organized as follows: We first review the related literature over the subject in Section 2. Then, we present the various features utilized for recognizing actions in still images, especially the MIL approach for objects in 3. In Section 4, we present the extensive evaluation of the features in the Stanford 40 actions [1] dataset. Section 5 finalizes the discussion with the conclusions and possible future directions.

2 Related Work

Human action recognition has been an active research area for computer vision for a while. For an extensive review, the interested reader can refer to one of the recent surveys over the subject [6, 7] and the references therein. Most of the existing work focuses on action recognition in videos, which makes use of motion cues and temporal information [8]. Action recognition in still images, however, is a more challenging problem, due to the lack of motion information and the difficulty of foreground subject segmentation.

In comparison to the large amount of work available for action recognition in videos, action recognition in still images is a less studied problem and is recently gaining attention. Wang, et al. [9] utilize deformable template matching for computing the distance between human poses and grouping similar poses. Thurau and Hlavac [10] use non-negative matrix factorization on pose primitives, where the pose primitives are learnt from non-cluttered videos and applied to images for finding the closest pose. In [11], the pose models are learnt from action images and those models are applied to classify actions in videos.

In more recent work, Yao and Fei Fei [12] have looked into the relationship between poses and objects and model the interactions using grouplet features. Object-person interactions are explored in other works such as [13, 2, 3, 14]. Delaitre et al. [15] has studied the use of bag-of-features and part-based representations using structural SVMs. Later on, Yao et al. [16] explore the use of random forests with discriminative decision trees. In their most recent work, Yao et al. [1] propose a part and attribute based model, which makes use of explicit object detectors for aiding action recognition in still images.

Prest et al. [4] also propose weakly supervised learning of human-object interactions. In [4], the objects having similar relative location with respect to the person are searched for the most recurring configuration for each action. For each image, their formulation is restricted to select one object window, whereas in our MIL approach, more than one object region can contribute to the recognition of the actions. Moreover, we do not enforce any spatial constraint for the objects and allow contributing object windows to come from any region of the image.

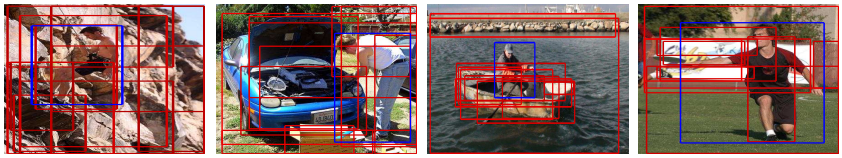

Fig. 1. Candidate object regions found by objectness measure [5]. The person bounding box is shown in blue and object regions are in red. Candidate object regions form the instances of the corresponding MIL bags.

3 Multiple Features for Actions in Still Images

3.1 Multiple Instance Learning for Candidate Object Regions

In order to recognize actions in still images, the related objects can be particularly important. In this paper, instead of using explicit object detectors, we investigate whether we can automatically learn potential object regions that can boost action recognition performance. For this reason, we extract several candidate object regions and use these object regions in a Multiple Instance Learning(MIL) framework.

We assume that the objects that the people are interacting with are visually salient objects. We use objectness measure [5] for finding visually salient regions within the image. Objectness measure uses several cues (such as multi-scale saliency, color contrast, edge density, etc.) in an image to identify regions for generic objects. We use this measure to identify candidate object hypotheses. Figure 1 shows example images. As it can be seen, in some images, objectness measure is able to locate objects of interest such as rowing boat. However, this measure also generates some noisy regions that do not include any related object.

In our implementation, we sample 100 windows from each image based on their objectness measure, i.e, the probability of containing an object. The authors of [5] recommend sampling 1000 image windows to cover all possible objects, but it would be very costly for the scalability of the approach. Therefore, we limit the sampling to 100 windows. We then extract dense SIFT feature vectors from each of these windows, and describe each via its bag-of-words representation using $2 \times 2 + 1 \times 1$ spatial tiling. The used codebook size is 1000 and the final feature vector dimensionality is 5000.

After sampling 100 windows from each image, we use k-means over the appearance feature vectors and group these 100 windows into 10 clusters. We use the cluster centers as our representation of candidate object regions. This step reduces the number of candidate object regions and also focuses on more condensed regions of potential objects. It is also likely that this clustering step smooths out the effect of the noise within candidate object regions.

As a result, we obtain multiple candidate regions from each image, some of which are likely to contain relevant objects for particular actions. However, we do not know which of these regions are related to the action. This case is particularly suitable for Multiple Instance Learning (MIL), since there are several candidate

regions where some of them are noisy and some of them could potentially include related contextual object for the action. In the traditional supervised learning, the learning procedure works over instances x_i and their corresponding labels y_i. In contrast, multiple instance learning operates over bags of instances, where each bag B_i is composed of multiple instances x_{ij}. In our formulation, each image can be considered as a "*bag*" of possible object regions and each extracted candidate object region is a corresponding "*instance*" inside the bag. A bag B_i is labeled as positive, if at least one of the instances x_{ij} within the bag is known to be positive, whereas it is labeled as negative, if all the instances are known to be negative. This form of learning is referred as "semi-supervised" (or "weakly supervised"), since the labels for the individual instances (in our case, individual object regions) are not available, and only labels of the bags are given.

Given the extracted candidate bounding boxes, we adopt Multiple Instance Learning with Instance Selection (MILES) [17] algorithm for learning the related object regions. MILES algorithm works by embedding the original feature space x, to the instance domain $\mathbf{m}(B)$. Each bag corresponds to an image and therefore has an associated label $Y_i \in A$, where $A = \{a_1, ..., a_M\}$ is the possible set of M actions. Each bag is represented by its similarity to each of the instances in the dataset. In our formulation, since the number of images and number of windows extracted from each image is high, we can cluster the instances and find the "concept instances" for a more scalable representation. The similarity between bag $\mathbf{B}_i$ and a concept instance c_l is defined as

$$s(c_l, \mathbf{B}_i) = \max_j \exp\left(-\frac{D(x_{ij}, c_l)}{\sigma}\right), \tag{1}$$

where $D(x_{ij}, c_l)$ measures the distance between a concept instance c_l and a bag instance x_{ij} and σ is the bandwidth parameter. We use the Euclidean distance for D and for the concept instances c_l, we either use all the object regions or cluster the instances via k-means and use the cluster centers as c_l for each action. We evaluate the effect of this clustering in the experiments section.

Each bag can then be represented in terms of its similarities to each of these target concepts and this mapped representation $\mathbf{m}(B_i)$ can be written as

$$\mathbf{m}(B_i) = [s(c_1, B_i), s(c_2, B_i), \ldots, s(c_N, B_i)]^T. \tag{2}$$

Using this embedded representation, we then train an L2-regularized SVM with RBF kernel for each action class in a one-vs-all manner.

3.2 Facial Features for Action Recognition

For quite a number of actions, facial features can be an indicator of the ongoing action. For example, for catching action, the person can be looking into some direction focusing on the thrown object. Similarly, the objects around the face can be a cue for the actions such as talking on the phone, brushing teeth, and so on. Based on this observation, we investigate the effect of facial features for

Fig. 2. The first three images show the person bounding boxes and the face detector outputs, and the latter ones shows face regions determined wrt person bounding boxes.

generic action recognition in still images. In [18], it has been shown that facial features can be useful in interaction recognition, and here we investigate their effect to generic actions.

With this intuition, we run a face detector [19] and for images in which the faces are detected, we extract an extended bounding box around the face area as shown in Fig. 2. For the images in which no face is detected, we use the top region of the person bounding box as the face area. From these regions, we extract dense SIFT [20] features and employ bag-of-words. We cluster the face images and form a codebook using k-means ($k = 1000$). Then using 2×2 spatial tiling, we extract the codeword histograms from each of the spatial bins. We also concatenate the bag-of-words histogram of the overall face region, hence the final feature vector size becomes 5000.

3.3 Additional Features

We also include additional features which are frequently used for action recognition to our evaluation framework. For this purpose, we extract the Histogram of Oriented Gradient(HOG) features from the person regions in the image. Furthermore, bag-of-words(BoW) representations extracted from person bounding boxes have also been evaluated. For this purpose, similar to BoW extracted around the faces, the SIFT features are densely extracted from the person regions and k-means clustering (with $k = 1000$) is applied to form the corresponding codebook. Then, 3×3 spatial binning is applied and all the codebook histograms from each spatial bin are concatenated with the global histogram extracted from the whole person region. In the end, the final feature vector for person BoW representation is 10000 dimensional.

In addition to the features extracted from the person region, we also consider the features from the original image and form the BoW representation from the whole image. This is also extracted in a similar manner to person BoW, where $3 \times 3 + 1 \times 1$ spatial tiling is used and the resulting feature vectors from each spatial bin are concatenated altogether to form a 10000-dimensional vector.

4 Experiments

4.1 Datasets and Experimental Setup

In the experiments, we use the Stanford 40 Actions dataset [1], which contains 40 actions and 180-300 images for each action. We use the same train/test split

Fig. 3. An example execution of the MIL framework (best viewed in color). Amongst the 10 example object regions extracted by [5] from the training set, the top 3 regions that contribute to the classification are shown in green, cyan and blue respectively.

provided, which includes 4000 train images and 5532 test images. The bounding boxes for the people doing the action are provided with the dataset. In our experiments, we use these bounding boxes in extracting person/face HoG and BoW features, both in the train and test phases, simulating the case with a perfect person detector, as in [15].

We train a one-vs-all SVM classifier for each of the feature representations separately. The final classification scores are obtained by linearly combining individual classifier confidences giving an equal weight for each feature representation.

4.2 Performance of the individual features

Example object/image regions that are discovered by the MIL training stage are shown in Fig. 3. For the visualization purposes, number of candidate object regions in this example run is limited to 10 and the top regions mapped to the most contributing concept instances are displayed. As it can be seen, the algorithm is quite successful in discovering the related object regions. In the "cooking" image, the dish region is discovered, whereas in "walking the dog" example, the dog is successfully located. The MIL method also finds the person region as a top contributing region in most of the cases.

In Table 1, we evaluate the effect of the clustering individual instances versus using all instances in the objectness-based MIL formulation. While the clustering provides a scalable representation that requires much less time (clustering with $k = 300$ runs ~ 14 times faster than no clustering case), using all the candidate object regions for instance embedding produces far more effective results in terms of the classification performance.

We then evaluate the performance of the individual features. Accuracy and mean Average Precision(mAP) values achieved by using individual features are

Table 1. Accuracy and mean average precision(mAP) achieved by our MIL approach.

	accuracy	mAP
objectMIL ($k = 300$)	37.08	34.03
objectMIL ($k = 1000$)	46.78	46.01
objectMIL (no clustering)	**51.34**	**51.80**

Table 2. Accuracy and mean average precision(mAP) of individual features and the combinations.

	accuracy	mAP
personHOG	24.75	19.35
personBoW	28.56	21.53
faceHOG	14.01	10.37
faceBoW	17.93	13.83
imgBoW	33.51	26.32
objectMIL	51.34	51.80
imgBoW+objectMIL	52.30	52.23
All(w/o objectMIL)	41.47	36.63
All	**55.93**	**55.55**
Yao [1]	NA	45.7

shown in Table 2. As it can be seen, the best performance is obtained using our MIL framework over the candidate object regions. This demonstrates that without explicit object detectors, we can extract useful information from the candidate object regions generated, in a weakly supervised manner by means of the multiple instance learning formulation.

Person-based features are also informative. Interestingly, performance of the BoW extracted from the whole image is higher than BoW extracted from the person bounding boxes only. This indicates that, the overall image contains more information than the person bounding box itself and the context information accompanying the person is useful for action recognition.

Figure 4 shows the performance of the individual features with respect to each action. Overall, the combination of all the features works the best for most of the actions. Interestingly, for some actions such as "climbing, rowing a boat, smoking and using computer" the performance of the proposed MIL framework performs better than using all features. BoW features over the facial region works best for the actions like "climbing, rowing a boat, playing violin, jumping, watching TV, shooting an arrow, brushing teeth". This is not surprising, since in these actions either the facial expression is representative of the action or the related object is closer to the face area. For "climbing, riding a horse, rowing a boat, playing guitar, riding a bike, playing violin, jumping, throwing frisby, running, applauding, holding an umbrella" kind of actions, HoG features around the face area are even more informative than the BoW counterpart. This may be due to the importance of orientation of faces in these type of actions.

4.3 Comparison to state-of-the-art

We compare our method to the state-of-the-art method of Yao et al [1] in Table 2 and Figure 5. Yao et al.'s method is based on part and attribute representation, where each image is represented via a sparse set of "*action bases*". These action bases are defined as the high level interactions between individual action attributes and action parts. In this respect, the attributes that describe an action

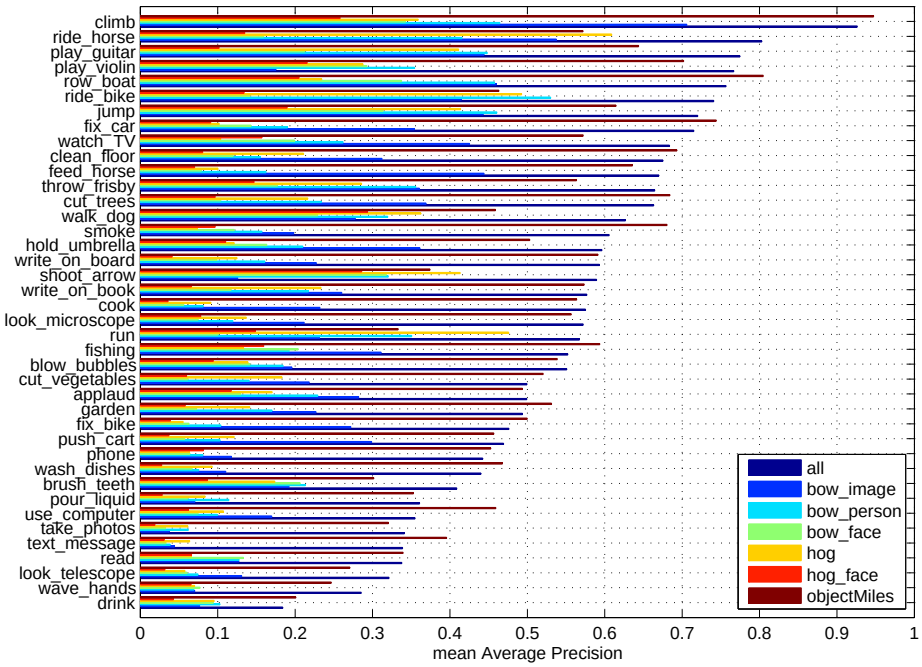

Fig. 4. Per action mAPs for each of the features (best viewed in color and magnified). Overall, combining all the features' responses works the best. For some actions, the performance of object MIL approach is even better than the combination.

are annotated and a discriminative binary classifier is trained for each action attribute. Moreover, each part is modeled by the output of an object detector (pre-trained on ImageNet data) or a pre-trained poselet detector [21].

In Table 2, `imgBoW+objectMIL` result shows the performance of our method without using any person bounding box information and `All` shows the performance of the proposed method using all features described in Section 3. Compared to the state-of-the-art result of Yao et. al [1], our method achieves significantly better results, while using much less supervision. Even without assuming the availability of a person detector, the objectness-based MIL method combined with image BoW features provide $\sim 6.5\%$ performance improvement in this extensive dataset.

Looking at Fig. 5, we observe that our method outperforms the parts and attributes method of [1] for most of the actions, especially for "climbing, playing guitar, playing violin, fixing a car, cooking, smoking, cooking, applauding, phoning, taking photos, texting message" actions. This indicates that without using any explicit object/part detector, our method is able to discover the recurring objects or image regions that contribute to the recognition. On the contrary, [1] outperforms our method especially in "riding a horse, rowing a boat, riding a bike, walking the dog, shooting an arrow, fishing, holding an umbrella, running"

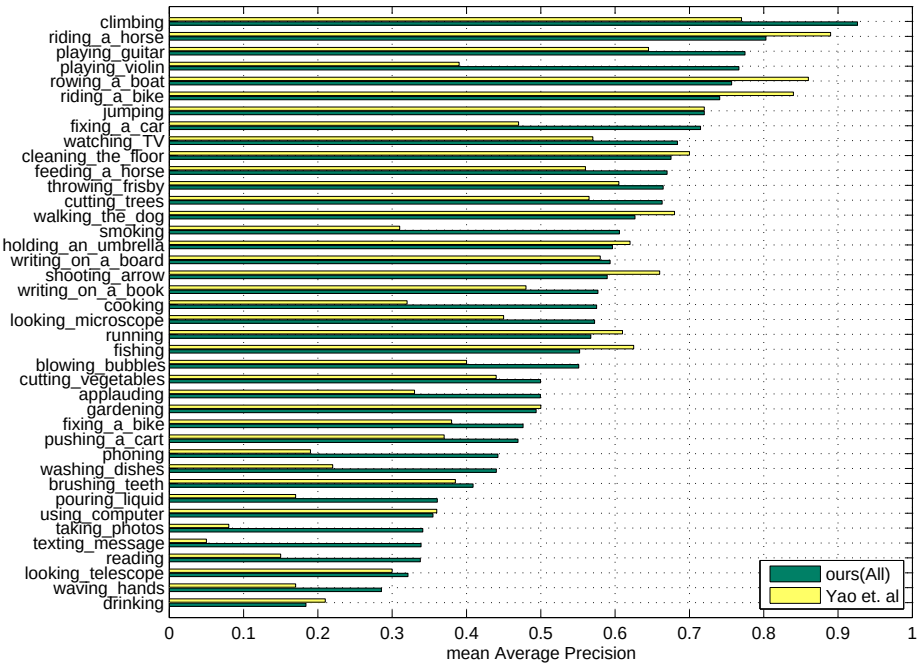

Fig. 5. Comparison of the proposed approach with that of Yao et al. [1] in terms of classification performance of the individual action classes.

actions. This may be due to the success of the explicit detectors in locating certain objects and also due to the shared nature of the attribute classifiers.

5 Conclusions and Discussion

In this paper, we have proposed a method that leverages the candidate object regions in a weakly unsupervised manner via Multiple Instance Learning and evaluated the performance of this method in combination with other visual features for human action recognition in still images. Our experimental results show that the proposed MIL framework is suitable for extracting the relevant object information, without the need for explicit object detectors. We have achieved better classification performance compared to the state-of-the-art on the extensive Stanford 40 actions still image dataset.

Our findings indicate possible future directions, particularly, using richer representations over salient object regions and improving weakly supervised learning of relevant objects.

Acknowledgments. This work was supported by a Google Research Award.

References

1. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L.J., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: International Conference on Computer Vision (ICCV), Barcelona, Spain (November 2011)
2. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. TPAMI **31** (2009) 1775–1789
3. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR, San Francisco, CA (June 2010)
4. Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. IEEE TPAMI **34** (2012) 601–614
5. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: IEEE Conf. on Computer Vision and Pattern Recognition, San Francisco, USA (2010)
6. Poppe, R.: A survey on vision-based human action recognition. Image Vision Computing **28** (June 2010) 976–990
7. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. CVIU **115** (February 2011) 224–241
8. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008)
9. Wang, Y., Jiang, H., Drew, M.S., Li, Z.N., Mori, G.: Unsupervised discovery of action classes. In: CVPR. (2006)
10. Thurau, C., Hlavac, V.: Pose primitive based human action recognition in videos or still images. In: CVPR. (2008)
11. Ikizler-Cinbis, N., Cinbis, R.G., Sclaroff, S.: Learning actions from the web. In: Int. Conf. on Computer Vision. (2009)
12. Yao, B., Fei-Fei, L.: Grouplet: a structured image representation for recognizing human and object interactions. In: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA (June 2010)
13. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for static human-object interactions. In: Workshop on Structured Models in Computer Vision. (2010)
14. Delaitre, V., Sivic, J., Laptev, I.: Learning person-object interactions for action recognition in still images. In: NIPS. (2011)
15. Delaitre, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: BMVC. (2010)
16. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In: CVPR, Springs, USA (June 2011)
17. Chen, Y., Bi, J., Wang, J.Z.: Miles: Multiple-instance learning via embedded instance selection. IEEE TPAMI **28** (2006) 1931–1947
18. Patron-Perez, A., Marszalek, M., Reid, I., Zisserman, A.: High five: Recognising human interactions in tv shows. In: British Machine Vision Conference. (2010)
19. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR. (2001)
20. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60** (2004) 91–110
21. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV. (2009)