Recognizing Actions from Still Images

Nazli Ikizler, R. Gokberk Cinbis, Selen Pehlivan and Pinar Duygulu
Bilkent University, Dept of Computer Engineering, 06800, Ankara, Turkey
{*inazli, cinbis, pselen, duygulu*}*@bilkent.edu.tr*

Abstract

In this paper, we approach the problem of understanding human actions from still images. Our method involves representing the pose with a spatial and orientational histogramming of rectangular regions on a parse probability map. We use LDA to obtain a more compact and discriminative feature representation and binary SVMs for classification. Our results over a new dataset collected for this problem show that by using a rectangle histogramming approach, we can discriminate actions to a great extent. We also show how we can use this approach in an unsupervised setting. To our best knowledge, this is one of the first studies that try to recognize actions within still images. [1]

1. Introduction

Long before the evolution of the video technology, the human actions were conveyed via static images. The newspapers still use action photography to picturize their news. Although motion is a very important cue for recognizing actions, when we look at such images, we can more or less understand human actions in the picture. This is mostly true in news or sports photographs, where the people are in stylized poses that reflect an action. Figure 1 shows some example images. However, understanding human actions from still images is a widely ignored problem of computer vision.

In this paper, we try to address this problem and answer the question of "Can we recognize human actions within a single image?". This problem is considerably harder than classical object recognition since there is high amount of articulation. We need shape descriptors that are able to model the variations caused by high articulations. Our approach starts with employing a pose extractor, and then representing the pose via distribution of its rectangular regions. By using classification and feature reduction techniques, we test our representation via supervised and unsupervised settings.

Figure 1. Actions in still images.

2. Related Work

Most of the effort on understanding the human actions involves video analysis with fundamental applications such as surveillance and human computer interaction. In particular, a number of approaches are proposed for recognizing actions over video sequences(see [3, 4] for extensive reviews). However, action recognition on single images is a mostly ignored area. This is due to various challenges of this topic. The lack of region model in a single image precludes discrimination of foreground and background objects. The presence of articulation makes the problem much harder, for there is a large number of alternatives for human body configuration. Thus, the problem of action recognition on still images becomes a challenging problem.

Recognition of actions from still images starts with finding the person inside the image and inferring the pose of it. There are many studies in finding person images [6],localizing the persons in still images [1], or pedestrian detection [10]. Dalal and Triggs propose a very successful edge and gradient based descriptor, called Histogram of Oriented Gradients [2]. Zhu *et al.* advances HOG descriptors by integrating HOG and AdaBoost to select the most suitable block for detection [13]. Oncel *et al.* [11] define a covariance descriptor for human detection.

For inferring the human pose from 2D images, there are a few recent studies. Ramanan et al. presents an iterative parsing process for pose estimation of articu-

[1]This research is partially supported by TUBITAK Career grant 104E065 and grants 104E077 and 105E065.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

lated objects [7]. Ren *et al.* presents a framework for detecting and recovering human body configuration [9].

Wang *et al.* also partially addresses the problem of action recognition using single images [12]. They represent the overall shape as a collection of edges obtained through canny edge detection and propose a deformable matching method to measure distance of a pair of images. However, they only tackle the problem in an unsupervised manner and within single sports scenes.

3. Our Approach

In still images, understanding motion is not a straightforward process. In the presence of motion, it is relatively easier to localize the person, whereas, in still images, we need to estimate the place and pose of the person. However, in the presence of background clutter and occlusions, it is not very straigtforward to localize the person and represent the pose.

3.1. Pose extraction from still images

We first use the method of Ramanan [7] to extract a pose from the still image. The approach uses edge and region features, and constructs two deformable models using Conditional Random Fields (CRF). Edge-based deformable model consists of K number of parts denoted as l_i. Using the part information, the configuration of the model is represented as $L = [l_1, l_2, \ldots l_k]$. This representation is a tree structure, and each part corresponding to a node of the tree has a single parent. The deformable model equation is defined as follows

$$P(L|I)\alpha \exp(\sum_{i,j \in E} \Psi(l_i - l_j) + \sum_i \phi(l_i))$$

Here, $\Psi(l_i - l_j)$, is the priori information of relative arrangements of part i with respect to its parent part j. In the study, the shape prior expresses in terms of discrete binning. $\phi(l_i)$ corresponds to local image features extracted from the oriented image patch located at l_i. The overall edge-based deformable model is used to estimate the initial body part positions. Then, using the previously obtained estimate, the method creates a region model(parse) that represents an image for each one of the body parts. Then, information obtained from part histograms become the ground for the region-based deformable model. The initial estimates of body positions from region-base model are utilized to build a second region-based model. The procedure continues iteratively by constructing a region model that is based on color evidence.

While pose extraction is still in its infancy, it gives some idea about the overall posture of the person. Figure 2 shows example images and their corresponding poses. We use these initial parses as basis and extract silhouettes by thresholding over the probability maps.

Figure 2. Pose and rectangle extraction.

3.2. Representing the pose

For describing the human pose, we make use of rectangular patches that we have initially introduced in [5]. These patches are extracted in the following way: Given the human silhouettes, we search for rectangular regions over this silhouette using convolution of a rectangular filter on different orientations and scales. We use undirected rectangular filters, following [8]. The search is performed using 12 tilting angles, which are 15° apart. To tolerate the differences in the limb sizes and in the varying camera distances to the subject, we perform the rectangle convolution over multiple scales.

After finding rectangles over the silhouettes, we use Histogram of Oriented Rectangles(HORs [5]) for representing the pose. We compute the histogram of extracted rectangular regions based on their orientations. The rectangles are histogrammed over 15° orientations. For still images, we do this histogramming over the spatial circular grids and define circular HORs (CHORs), as opposed to original $N \times N$ grid form. This is mostly because we don't know the explicit height of the human figure due to the discrepancies of the parse. Using circular grid helps us to capture the angular positions of the parts more reliably in this case. We use the center of the highest probability region of the parse as the center of our circular grid. The bins of this circular histogram are 30° apart, making 12 bins in total. We depict this process in Fig 3.

4. Action recognition

As we discuss in [5], HORs are quite compact representations for action recognition in videos. However, we can further densify the representation. In fact, in still image case, we have much less examples for action classes, therefore feature reduction is necessary

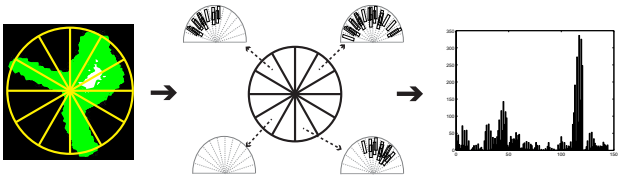

Figure 3. Pose representation using CHORs. Circular grid is centered to the maximum value of the probability parse.

for learning. For this purpose, we first apply Linear Discriminant Analysis(LDA) in our feature space. By using LDA, we reduce the feature dimension from 144 to 50.

We then train one-vs-all SVM classifiers for each action separately and use the highest probable class label. We form the SVM classifiers using rbf kernels.

For evaluating the performance of our method on unsupervised classification, we also apply clustering, and make a qualitative evaluation of clusters. We run kmeans over the data for 100 times, and take the clustering that minimize the intra-cluster distance and maximize the inter-cluster distance. The respective results are given in Section 5.

5. Experimental Results

Datasets: For recognition of actions from still images, we collected a dataset from various sources like Google Image Search, Flickr, BBC Motion database, etc. This dataset consists of 467 images and includes six different actions; these are running, walking, catching, throwing, crouching and kicking. We choose this subset of actions, because these are mostly visually identifiable actions from single images. Example images for each action is shown in Fig 1. This image collection involve a huge amount of diversity by means of viewpoints, shooting conditions, cluttered backgrounds, resolution. We apply leave-one-out cross-validation and report the results.

We also test our descriptor's performance for the case of unsupervised classification. For this purpose, we used Wang *et al.*'s skating images dataset [12]. This dataset is a collection of 1432 images, where different figure skaters perform various moves.

Results: Our overall accuracy rate for supervised classification on still actions dataset is 85.1%. This is a surprisingly good result, given the fact that the images cover a wide range of poses (see Fig. 1) and foreground parses are not that perfect (Fig. 2). However, by using CHORs, these results show that we can still overcome most of such discrepancies and achieve high accuracy rates. Figure 5 shows examples for the correctly classified images by our approach. Note that the diversity of the images in the dataset is very large, with cluttered backgrounds, different poses, outfits and also carry items. We also present examples of the misclassified images in Fig. 6. It can be observed that some of the poses are very similar, indistinguishable even to the human eye, and also the lack of proper edge boundaries make the pose extraction harder. The corresponding confusion matrix for our representation with supervised classification is given in Fig. 7.

Figure 5. Examples for correctly classified images of actions running, walking, throwing, catching, crouching, kicking in consecutive lines.

We also present qualitative results of clustering with our approach. Figure 4 presents some of the clusters that we get with the Wang *et al.*'s dataset. We used $k = 100$ in our clustering execution. As seen, the clusters we get are quite coherent and each of them represents a certain pose.

6. Discussions and Conclusion

In this study, we present a novel method for action recognition from still images. Our method is based on

Figure 4. Clusters formed by our approach for the figure skating dataset

(a) catch,walk,catch,throw (b) run,run,run,kick

(c) catch,kick,walk,crouch (d) run,throw,run,run

(e) kick,walk,walk,catch (f) throw,walk,run,throw

Figure 6. Examples for misclassified images of actions running, walking, throwing, catching, crouching, kicking with their wrong classification labels.

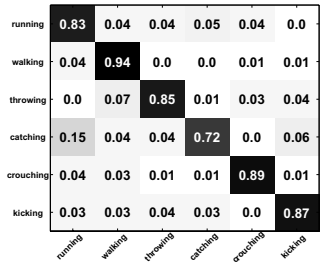

Figure 7. Confusion matrix for the still images action set

extracting parses of the human figure and representing it by means of spatial and orientational binning, using circular Histogram of Rectangles (CHORs). To our best knowledge, this is one of the very first efforts that try to discriminate actions within still images. Our high success rates over the challenging still action set shows that our method is quite competent in discriminating the actions.

References

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *PAMI*, 28(1), January 2006.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages I: 886–893, 2005.

[3] D. Forsyth, O.Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. Computational studies of human motion i: Tracking and animation. *Foundations and Trends in Computer Graphics and Vision*, 1(2/3):1–255, 2006.

[4] D. M. Gavrila. The visual analysis of human movement: A survey. *CVIU*, 73(1):82–98, 1999.

[5] N. Ikizler and P. Duygulu. Human action recognition using distribution of oriented rectangular patches. In *Human Motion Workshop LNCS 4814*, pages 271–284, 2007.

[6] S. Ioffe and D. Forsyth. Learning to find pictures of people. In *NIPS*, 1998.

[7] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006.

[8] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I: 271–278, 2005.

[9] X. Ren, A. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. *Proc. ICCV*, pages 824–831, 2005.

[10] D. Tran and D. Forsyth. Configuration estimates improve pedestrian finding. In *NIPS*, 2007.

[11] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *CVPR*, 2007.

[12] Y. Wang, H. Jiang, M. S. Drew, Z.-N. Li, and G. Mori. Unsupervised discovery of action classes. In *CVPR*, 2006.

[13] Q. Zhu, S. Avidan, M. Yeh, and K. Cheng. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. *CVPR*, 1(2):4, 2006.