We focus on the problem of still image-based human action recognition, which essentially involves making prediction by analyzing human poses and their interaction with objects in the scene. Besides image-level action labels (e.g., riding, phoning), during both training and testing stages, existing works usually require additional input of human bounding boxes to facilitate the characterization of the underlying human-object interactions. We argue that this additional input requirement might severely discourage potential applications and is not very necessary. To this end, a systematic approach was developed in this paper to address this challenging problem of minimum annotation efforts, i.e., to perform recognition in the presence of only image-level action labels in the training stage. Experimental results on three benchmark data sets demonstrate that compared with the state-of-the-art methods that have privileged access to additional human bounding-box annotations, our approach achieves comparable or even superior recognition accuracy using only action annotations in training. Interestingly, as a by-product in many cases, our approach is able to segment out the precise regions of underlying human-object interactions.