Action Detection in Crowd

Parthipan Siva
psiva@dcs.qmul.ac.uk
Tao Xiang
txiang@dcs.qmul.ac.uk

School of EECS,
Queen Mary University of London,
London E1 4NS, UK

As the amount of video content in use increases, methods for searching and monitoring video feeds become very important. The ability to automatically detect actions performed by people in busy natural environments is of particular interest for surveillance applications where cameras monitor large areas of interest looking for specific actions like fighting.

Action detection in crowd is closely related to the action recognition problem which has been studied intensively recently [1, 5, 6]. However, there are key differences between the two problems which make action detection a much harder problem. Specifically, recognition does not deal with localization of action amoung cluttered background. The existing action recognition methods are thus unsuitable for solving our problem. There have been a few attempts in the past three years on action detection [3, 4, 7, 8]. Nevertheless most of them are based on matching templates constructed from a single sample thus unable to cope with the large intra-class variations.

Motivated by the success of sliding-window based 2D object detection, we propose to tackle the problem by learning a discriminative classifier based on Support Vector Machine (SVM) from annotated 3D action cuboids and sliding 3D search windows for detection. To learn a discriminative action detector, a large number of 3D action cuboids have to be annotated, which is tedious and unreliable. To overcome this problem, we propose a novel greedy k nearest neighbour algorithm for automated annotation of positive training data, by which an action detector can be learned with only a single manually annotated training sequence.

To represent the action we use the trajectory based feature descriptor similar to that of Sun et al. [6], which has been shown to perform well in complex action recognition datasets such as the Hollywood dataset [5]. Trajectory based features encodes appearance and motion information of tracked 2D points. First, spatially salient points are extracted at each frame and tracked over consecutive frames, using 1-to-1 pairwise matching of SIFT descriptor, to form tracks $\{\mathbf{T}_1, \ldots, \mathbf{T}_j, \ldots, \mathbf{T}_N\}$. Second, the appearance information associated with each track $\mathbf{T}_j$ is represented by the average SIFT descriptor S_j. Third, the motion information of each track is represented using a trajectory transition descriptor (TTD) π_j.

The average SIFT S_j and motion descriptor π_j is then used to construct a bag of word (BoW) model for action representation with 1000 words for each type of descriptor. To take advantage of the spatial distribution of the tracks, we use all six spatial grids (1×1, 2×2, 3×3, $h3\times1$, $v3\times1$, and $o2\times2$) and the four temporal grids ($t1$, $t2$, $t3$, and $ot2$) proposed in [5]. The combination of the spatial and temporal grids gives us 24 channels each for S_j and π_j, yielding a total of 48 channels. We employ the greedy channel selection routine, similar to [5], to select at most 5 of the 48 channels because these 48 channels contain redundant information.

We model an action as a 3D spatio-temporal cuboid or window illustrated in Fig. 1(a) which we call the action cuboid. The action cuboid is represented by the multi-channel BoW histogram of all features contained within the cuboid. Our task is to slide a cuboid spatially and temporally over a video and locate all occurrences of the action in space and time (see Fig. 1(b)). To determine whether a candidate action cuboid contains the action of interest, a Support Vector Machine (SVM) classifier is learned.

Our training set consists of a set of video clips without the action $V^-_{i=1\ldots N_1}$, clips known to contain at least one instance of the action $V^+_{i=1\ldots N_2}$ and one clip which has been manually annotated by a cuboid $Q_{k=1}$ around the action. From V^-_i and V^+_i a set of spatially overlapping instances or cuboids $X^-_{i,j}$ and $X^+_{i,j}$ are extracted respectively, where $j = 1\ldots M_i$ are the number of cuboids in video i. Our objective is to automatically select positive cuboids/instances $X^+_{i,j=j^*}$ from each video i in the positive clip set V^+_i to add to the positive cuboid set Q_k which initially contains only the annotated positive cuboid $Q_{k=1}$. Unlike the traditional formulation of the multi-instance learning (MIL) problem here we have the assumption that a single positive instance has been annotated.

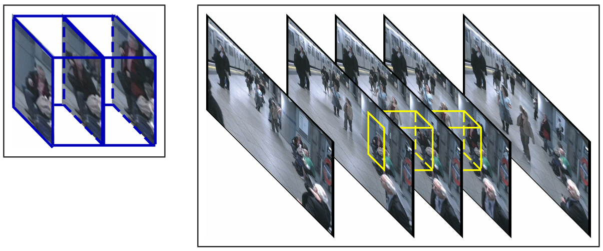

Figure 1: Action detection. (a) Action cuboid. (b) Action localized spatially and temporally.

Our solution is based on a greedy k nearest neighbour (kNN) algorithm. We iteratively grow the positive action cuboid set Q_k such that at each iteration we capture more of the intra-class variation and the kNN classifier becomes stronger. To that end, at each iteration we use the current kNN classifier to select one cuboid $X^+_{i^*,j^*}$ from the set of all available cuboids $X^+_{i,j}$ which is the closest to the action class represented by Q_k. To do this, (i^*, j^*) is selected as

$$\{i^*, j^*\} = \arg\min_{i,j}\left[\min_k d\left(X^+_{i,j}, Q_k\right) - \min_{l,m} d\left(X^+_{i,j}, X^-_{l,m}\right)\right], \qquad (1)$$

where $d(X,Y)$ is the distance function. After adding X_{i^*,j^*} to Q_k our kNN classifier is automatically updated. In order to reduce the risk of including false positive cuboids into our positive training set Q_k, we take a very conservative approach, that is from each positive clip we only select one positive cuboid.

Our method is validated on the CMU dataset [4] and a new dataset created from the iLIDS database [2] featuring unstaged actions performed under natural settings with large number of moving people co-existing in the scene. Our results demonstrate that the proposed action detector trained with minimal annotation can achieve comparable results to that learned with full annotation, and outperforms existing MIL methods.

[1] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:288–303, 2010. ISSN 0162-8828.

[2] HOSDB. Imagery library for intelligent detection systems (i-lids). In *IEEE Conf. on Crime and Security*, 2006.

[3] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T.S. Huang. Action Detection in Complex Scenes with Spatial and Temporal Ambiguities. In *IEEE International Conference on Computer Vision*, 2009.

[4] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *IEEE International Conference on Computer Vision*, October 2007.

[5] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[6] J. Sun, X. Wu, S. Yan, L.F. Cheong, T.S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2004–2011. IEEE, 2009.

[7] W. Yang, Y. Wang, and G. Mori. Efficient human action detection using a transferable distance function. In *Asian Conference on Computer Vision*, 2009.

[8] J.S. Yuan, Z.C. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2442–2449, 2009.