# Actions Recognition in Crowd Based on Coarse-to-Fine Multi-Object Tracking

Sixue Gong, Hu Han*, Shiguang Shan, and Xilin Chen

Key Lab of Intelligent Information Processing, Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
{sixue.gong,hu.han,shiguang.shan,xilin.chen}@vipl.ict.ac.cn

**Abstract.** Action recognition has wide applications from video surveillance, scene understanding to forensic investigation. While recent methods typically focus on a single action recognition from video clips, we investigate the problem of action recognition in crowd, which better replicates real video surveillance scenarios. We propose to perform actions recognition in crowd based on an efficient coarse-to-fine multi-object tracking algorithm. With Faster R-CNN as our human detector, we utilize a coarse-to-fine strategy for multi-object tracking in crowd, consisting of multi-object fast tracking and per-object fine tracking. The tracking results are used to extract the action cuboids, and spatial-temporal features are computed for action classification. We evaluate the proposed approach on a self-collected actions-in-crowd dataset, and two public domain databases (CMU and and MOT2015). The results show the effectiveness of the proposed approach for multi-action recognition in crowd.

## 1 Introduction

The recognition of both human and animal actions in videos plays a crucial role from automatic scene understanding to video surveillance. For example, in video surveillance, it is helpful to automatically discover suspects through action of interest detection. Also, a coach's workload can be alleviated if every players action statistics in a game can be automatically calculated. Considerable progress has been made in the past few years, particularly on action feature descriptors like Bag-of-Word (BoW) [1–3] and CNN features [4, 5]. These feature representation methods significantly improve the accuracy of action classifiers. Nevertheless, these approaches mainly focus on a single action recognition from videos. However, in many applications with various arising actions, such as video surveillance and group sports analysis, it is required to recognize multiple actions of interest (see Fig. 1). Most of the existing action recognition methods are not designed for recognizing multiple actions in crowd scenarios.

Multi-action recognition in crowd is non-trivial because of the challenges from designing human detection and tracking algorithms, robust action representations, and classification models. To bridge the gap between the applicability

* Corresponding author. E-mail: hanhu@ict.ac.cn (H. Han).


**Fig. 1.** In real applications, the scenarios often contain multiple actions in crowd: (a) actions including cycling and waving, and (b) actions including kicking, walking, and running. Multi-action recognition is necessary for exact scene understanding.

of existing action recognition methods and the needs of emerging applications, we propose an approach for multi-action recognition in crowd utilizing coarse-to-fine multi-object tracking. Given the individual subjects detected by a faster region-based convolutional neural networks (Faster R-CNN) [6], our tracking algorithm uses a linear quadratic estimation model, i.e., Kalman filter [7], for coarse but fast tracking of all the moving subjects, and a local sparse optical flow model for refined and stable tracking. Person-specific action tracks (cuboids) are then extracted based on the tracking results; each can be input to the traditional single action recognition algorithms for action recognition.

The main contributions of this work are as follows. (1) While existing action recognition approaches typically focus on a single action recognition per video clip, we explore a relatively underserved area of concurrent actions recognition per video clip; (2) We propose a coarse-to-fine multi-object tracking algorithm while balances the tracking speed and accuracy under crowd scenarios; (3) We also build an database to better replicate the action-in-crowd application scenarios.

## 2 Related Work

**Action Recognition.** There is a variety of work on action recognition from manually segmented video clips containing only one type of action [8, 5, 9–11, 1, 12–16]. The majority of these approaches aimed at utilizing trajectory [17–19] and spatial-temporal motion cues to get generative and discriminative features of actions. The early work on spatial-temporal feature extraction used smoothed and aggregated optical flow to model human motions [8]. Zhu at al. [13] used spatial-temporal interest points (STIP) as low-level motion features of the action segments, and used a multi-SVM classifier .

Jain at al. [10] trained a linear SVM classifier using the Divergence-Curl-Shear descriptor encoded by VLAD [20]. They employed the horizontal and vertical components of the flow field to compute the divergence, curl, and shear scalar values and consider all possible pairs of these kinematic features to capture more information through the joint distribution of the features. Ryoo at al. [11] categorized optical flows into multiple types based on their location and directions, and placed them into histogram bins as global features. These approaches mainly focused on designing an effective motion feature representation from a given video clip, but did not take into account the contextual information between the object and the background. However, such contextual information is of great importance in action recognition, particularly under scenarios with cluttered backgrounds (e.g., crowd). To address this issue, Gkioxari at al. [5] proposed a R-CNN method to perceive more visual information from a primary region of the subjects and a secondary region of the context areas.

The aforementioned approaches concentrate on action recognition with fixed background or with a single action inside [21]. To recognize actions in more complex scenes, Hu at al. [9] proposed a method that utilizes multi-instance learning (MIL) based SVM to handle the ambiguities in both spatial and temporal domains. They learned the SVM classifier from an action cuboid which is referred to as a bag containing more than one potential region and time slice. As for crowd scenes, Zhou at al [12] proposed a statistical framework to detect abnormal behaviors of the crowd scene by using a multi-observation hidden Markov model (HMM) of pedestrian trajectories. Such unusual action detection methods concern more about discovering abnormal behavior, but not for recognition of multiple concurrent actions. Siva at al. [1] integrated both static appearance features by bag-of-word (BoW), and motion features by trajectory transition descriptor (TTD), and used SVM in a sliding-cuboid manner to detect particular actions from a video.

**Visual Tracking.** Significant progress of visual tracking has been made these years. Tracking-by-detection [22–24] is one of the popular frameworks in visual tracking, in which the tracker is used to follow the object from frame to frame, and the detector is used to localize all Such tracking methods may be time-consuming if the object detector is complicated. Approaches like correlation filtering was proposed for more efficient object tracking [25–28], in which the correlation can be computed using Fast Fourier Transform (FFT), making it a fast tracker. In recent years, there are some approaches exploiting the rich feature hierarchies in CNNs for robust tracking [25, 29], which have shown state-of-the-art performance. As for multi-object tracking problem, Bae at al. [30] formulated a multi-object tracking problem based on the tracklet confidence, and used an incremental linear discriminant analysis for discriminating the appearances of objects to obtain reliable association between tracklets and detections. Xing at al. [31] proposed a two-stage framework to learn a tree-structured multi-view human detector to generate tracklets through particle filter in local stage.

## 3 Multi-action Recognition in Crowd

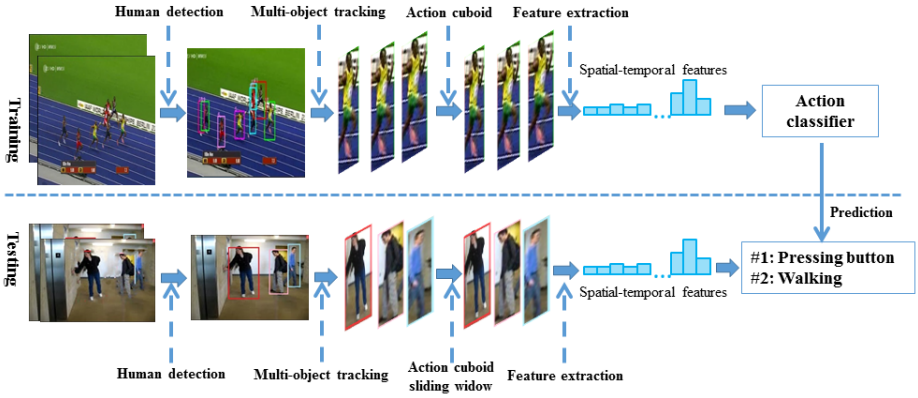



**Fig. 2.** Overview of the proposed approach for actions recognition in crowd.

Given a video clip $V$ in crowd with $M$ subjects (each has an individual action), the objective of our approach is to obtain an action label $l_i$ for each subject

$$L = \{l_i = C(F(N(T_i)))|T_i \in T(V)\}_{i=l}^{M}, \tag{1}$$

where $T(\cdot)$ is our coarse-to-fine multi-object tracker, which determines the candidate action tracks $T_i$ with actions of interest, and $T_i$ contains a series of bounding boxes $T_i = \{b_1, b_2, \cdots, b_{N_i}\}, i = 1, 2, \cdots, M$, where $b_j = [x_j, y_j, w_j, h_j]$ defines the bounding box's left-right location $(x_j, y_j)$ and size $(w_j, h_j)$; $N(\cdot)$ normalizes all the bounding boxes of each action track into the same size to form the 3D spatial-temporal cuboid; $F(\cdot)$ is to extract spatial-temporal features from each action cuboid $X_i = F(C_i), i = 1, 2, ..., M$; finally, the action recognition of one action cuboid is determined as

$$l_i = \arg\max_{\{1,2,...,M\}}(C(X_i)), \tag{2}$$

where $C(\cdot)$ calculates the confidences of all the action types that $X_i$ belongs to.

Traditional action recognition approaches often extract spatial-temporal features from the entire frames, resulting in features that are helpful for one dominant action recognition, but not for recognizing multiple concurrent actions. The proposed approach of multi-action recognition in crowd scenario is built upon multi-object tracking to address the above issue. An overview of the proposed approach can be seen in Fig. 2. We provide the implementation details of the proposed approach in the next section.

# 4 Implementation Details

## 4.1 Coarse-to-fine Multi-object Tracking

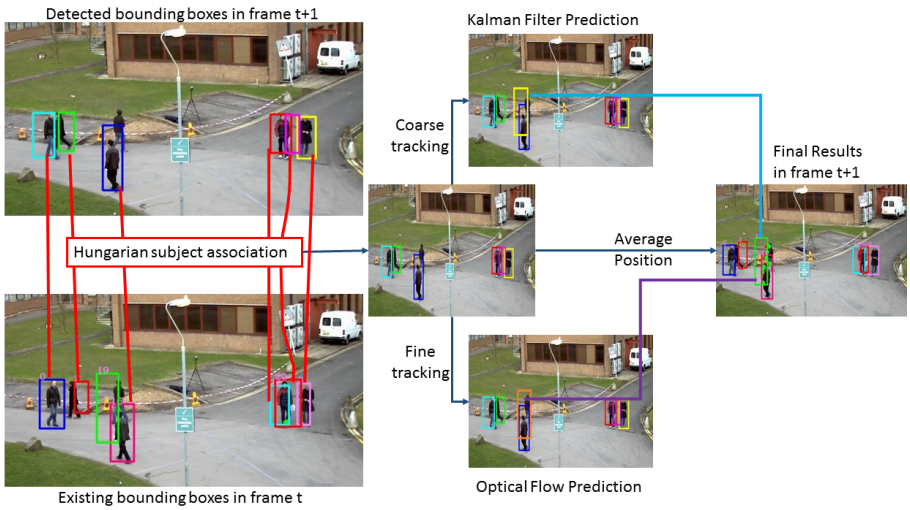



**Fig. 3.** A diagram of the proposed coarse-to-fine multi-object tracking algorithm.

The proposed coarse-to-fine multi-object tracking consisting of fast subject trajectory estimation via a linear quadratic estimation model, i.e., Kalman filter, and trajectory refinement using sparse optical flow. In particular, during our fast subject trajectory estimation, we assume that the moving object is in uniform velocity within a frame interval due to its short time span. If an object $O_i$ appears at frame $I_t$, the state of the object $O_i$ is denoted as $S_t^i = (p_t^i, s_t^i)$, where $p_t^i$ is the position, and $s_t^i$ is the size. Then, we construct the system state model of Kalman filter as

$$s(t) = \ddot{T}(t-1)s(t-1) + w(t), \tag{3}$$

where $s(t) = \begin{bmatrix} p_t^x \\ p_t^y \\ \Delta x_t \\ \Delta y_t \end{bmatrix}$, $s(t-1) = \begin{bmatrix} p_{t-1}^x \\ p_{t-1}^y \\ \Delta x_{t-1} \\ \Delta y_{t-1} \end{bmatrix}$; $\ddot{T}(t) = \begin{bmatrix} 1 & 0 & v_x & 0 \\ 0 & 1 & 0 & v_y \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ is the state transition matrix; w(t) is white Gaussian noise with zero mean.

Due to the unconstrained nature of crowd video surveillance scenario, objects may disappear due to occlusion and reappear at a later time, and new subjects may appear at any time. Thus, we need to detect new moving object every frame. Given the newly detected moving objects at frame $I_t$ and the estimated positions of existing objects at frame $I_t$, we apply the Hungarian algorithm [32] to calculate the optimal association matrix $A^* = \{a_i^*\}$ between subjects in frames

$I_{t-1}$ and $I_t$. In this process, each of the humans in $I_t$ is either associated to one of the existing tracklets and used to update the system state model or grouped as a new tracking target.

The fast tracking process is based on the assumption that the state posterior density of object motions follows a Gaussian distribution, which does not always hold in real tracking problems. As a result, abrupt appearance changes may occur in the candidate action tracks. To resolve this problem, sparse optical flow based on keypoints is employed to refine the fast tracking result. Based on the optimal assignment matrix $A^*$, if abrupt changes occur either in size or position of a subject, the sparse optical flow based tracking is applied to the subject. We compute the keypoints following the method in [33], and denote the computed keypoints at frame $I_t$ as $\{b^t_{i_1}, b^t_{i_2}, ..., b^t_{i_M}\}$. Then, the average displacement field of all keypoints between frames $I_t$ and $I_{t-1}$ can be calculated as

$$D^t_i = \frac{1}{M}\sum_{j=1}^{M} |b^t_{ij} - b^{t-1}_{ij}| \tag{4}$$

Accordingly, the state of object $O_i$ is updated as:

$$S^i_t = X^i_{t-1} + (D^i_t, 0) \tag{5}$$

Considering the success of Faster R-CNN [6] in a number of object detection tasks, we choose to use a Faster R-CNN [6] detector to detect the object of interest (persons) used by our tracking algorithm. In particular, we use a VGG-M model [34] with 5 convolutional (conv) layers in Faster R-CNN, and train it on PASCAL VOC 2007 dataset [35] for human detection. Such a detector takes about 70ms to process one frame (resized to $1000 \times 600$ inside Faster R-CNN) on a Titan X GPU. Our coarse-to-fine tracking runs really fast, about 30ms per frame ($640 \times 480$) on an Intel i7 3.6GHz CPU. Thus, the total detect and tracking time remains about 100ms.

### 4.2 Action Recognition

Our coarse-to-fine tracker determines a number of candidate action tracks $T_i$ from a video; each contains one subject if there is no tracking error. However, the number of actions in each track, and the action's spatial and temporal locations are still not known. Thus, the state-of-the-art action recognition methods designed for single action action recognition still cannot be applied directly. Another problem is that the sizes of the successive bounding boxes in $T_i$ may differ from each other because of the subject appearance changes in scale and pose (also known as the heterogeneous issue in object recognition [36–38]). Thus, before extracting action informative features, we first resize all the bounding boxes in $T_i$ into the same size of $20 \times 35$. We then do cuboid sliding inside the normalized $T_i$ temporally, and each cuboid is used for action feature extraction. Following the settings in [1], we also restrict the length of acceptable cuboids between $L_{min}$=20 and $L_{max}$=75 frames. If $N_i$ (the frame number of $T_i$) is larger than

$L_{max}$, each sliding cuboid has 75 frames; if $L_{min} \leq N_i \leq L_{max}$, each sliding cuboid has $N_i$ frames; otherwise, action track $T_i$ is deserted directly and will not be used for action recognition.

Each sliding 3D spatial-temporal cuboid is supposed to have one dominant action of one subject, and thus single action recognition methods can be used for the final action recognition. Without loss of generality, we use two typical action feature extraction methods: Biologically Inspired Feature (BIF) [39] and bag-of-words (BoW) [1]. In its simplest form, the extraction of BIF consists of two layers of computational units, where simple $S_1$ units in the first layer are followed by complex $C_1$ units in the second layer. The $S_1$ units correspond to the classical simple cells in the primary visual cortex [40]. They are typically implemented with the convolution of a preprocessed image $\mathbf{\Gamma}$ with a family of Gabor filters [41],

$$\psi_{u,v}(\mathbf{z}) = \frac{\|\mathbf{k}_{u,v}\|^2}{\sigma^2} e^{-\frac{\|\mathbf{k}_{u,v}\|^2 \|\mathbf{z}\|^2}{2\sigma^2}} \left[ e^{i\mathbf{k}_{u,v}\mathbf{z}} - e^{-\frac{\sigma^2}{2}} \right], \tag{6}$$

where $\mathbf{z} = (x, y)$, $\sigma$ is the relative width of the Gaussian envelope function w.r.t. the wavelength, and $u$ and $v$ are the orientation and scale parameters of Gabor kernels, respectively. The wave vector $\mathbf{k}_{u,v}$ is defined as,

$$\mathbf{k}_{u,v} = k_v e^{i\phi_u}, \tag{7}$$

with $k_v = \frac{k_{max}}{f^v}$ defining the frequency, and $\phi_u = \frac{\pi u}{8}$ defining the orientation. $k_{max}$ and $f$ are constants specifying the maximum frequency and scaling factor between two neighboring kernels, respectively. The $C_1$ units correspond to cortical complex cells which are robust to shift and scale variations. They can be calculated by pooling over the preceding $S_1$ units with the same orientation but at two successive scales. To compute $S_1$ layer features, we build a family of Gabor filters similar to those in [41], but we use 8 orientations and 12 scales. We apply "MAX" pooling operator and "STD" normalization operator to extract $C_1$ features from the $S_1$ layer. The $S_1$ layer provides a multi-scale representation for face images, and the $C_1$ layer provides robustness against translation, rotation, and scaling changes of the subjects.

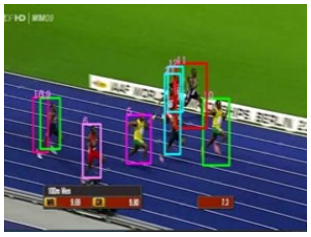
The calculation of our BoW feature is the same as the one used in [1], which was based on SIFT appearance descriptor. After the feature extractions, we used a SVM with an RBF kernel as the action classifier.

## 5 Experimental Results

### 5.1 Settings

During the training of the Faster R-CNN human detector, we set the learning rate base_lr as 0.01, and use a step learning rate policy with a weight of 0.1 every 40K iterations (and a maximum iteration number of 100K). The batch_size and iter_size were both set to 1. For divisions of the training and testing sets on different databases, we randomly use half clips of each scenario for training, and the remaining for testing.

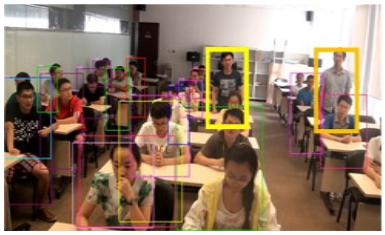
(a) Walking of multiple subjects



(b) Running of multiple subjects

(c) Kicking and running

(d) Sitting and one standing



(e) Sitting and two standings

**Fig. 4.** Examples of the video clips from the multi-action in crowd dataset we collected.

**Table 1.** Comparisons (in MAPs) with the baseline action recognition methods on the multi-action in crowd dataset we collected under scenarios (a–e).

| Scenarios | (a) | (b) | (c) | (d) | (e) |
|---|---|---|---|---|---|
| BIF | 0.29 | 0.43 | 0.14 | 0.67 | 0.55 |
| BoW | 0.26 | 0.34 | 0.17 | 0.52 | 0.50 |
| Proposed (BIF) | 0.36 | 0.51 | 0.23 | 0.65 | 0.61 |
| Proposed (BoW) | 0.32 | 0.40 | 0.25 | 0.51 | 0.56 |

### 5.2 Multi-action Recognition in Crowd

**Our dataset.** Most crowd datasets such as KTH, YouTube, Hollywood2, and UIUC [42–45] are focusing one action recognition of the entire crowd, not for the purpose of multiple concurrent actions recognition. Therefore, we build a multi-action dataset based on public domain databases and videos from the Internet. As shown in Fig. 4, there are five typical scenarios in the collected dataset: Fig. 4 (a) is walking of multiple subjects, Fig. 4 (b) is running of multiple subjects, Fig. 4 (c) is kicking and running, Fig. 4 (d) is one sitting and one standing, Fig. 4 (e) is one sitting and two standings. For each video clip, we manually provided the ground-truth bounding boxes and action labels of individual subjects.

We tested the performance of two baseline approaches using BIF and BoW, respectively, and their performance under the proposed framework (denoted as Proposed (BIF) and Proposed (BoW), respectively). We report the mean average precision (MAP) for each action category. When a video clip contains more than


**Fig. 5.** An example of the multi-action recognition results by the proposed approach on the multi-action dataset we collected.

**Table 2.** Comparisons (in MAPs) with the baseline on the CMU dataset.

| | Jump-Jacks | Pick-up | Push button | 1-hand wave | 2-hand wave |
|---|---|---|---|---|---|
| QMUL[1] | 0.36 | 0.68 | 0.94 | 0.45 | 0.54 |
| Proposed (BIF) | 0.36 | 0.51 | 0.23 | 0.65 | 0.61 |
| Proposed (BoW) | 0.32 | 0.40 | 0.25 | 0.51 | 0.56 |

one actions or the same actions but of more than one subjects. The MAP is the average of these multiple actions or subjects.

As shown in Table 1, in the scenarios when there are more than one actions, e.g., scenarios (a,b,c,e), the proposed approach using BIF and BoW features significantly outperforms the baseline approach. This is understandable because these baseline methods does not explicitly use multi-object tracking, and are difficult in handling multi-actions. For scenario (e) with a single action, the baseline approaches making use of contextual information, still work better than our method. In cropped action cuboids separate the multiple actions into separate ones but each losses the contextual information which is also helpful for action recognition.

**CMU dataset.** The CMU dataset [46] contains some crowd scenes, but most of the videos contain one a dominant action of one subject. So this database does NOT represent the scenarios that this work is focusing on. However, since there are not known multi-action in crowd databases, we still provide the results on this database as a reference. We use the state of the art results by QMUL method [1] as the baseline performance on CMU.

Not surprisingly, as shown in Table 2, the baseline methods work reasonably well under most scenarios except for the two-hand wave scenario. Such an observation actually provides a strong support of the usefulness of the proposed approach in recognizing multiple different actions or the same actions by multiple individuals.

### 5.3 Evaluation of the Tracking Module

Since the tracking module in the proposed approach plays an important role, we evaluate the tracking algorithm on the pubic database MOT2015 [47] using 9 video sequences (TUD-Stadtmitte, TUD-Campus, PETS09- S2L1, ETH-Bahnhof, ETH-Sunnyday, ADL-Rundle-6, ADL-Rundle-8, KITTI-17 and Venice-

**Table 3.** Evaluation of the tracking module in the proposed approach on MOT2015.

| Measures | Recall | Precision | MOTA | MOTP |
|---|---|---|---|---|
| Coarse tracking | 71.1 | 28.3 | 32.3 | 69.6 |
| Coarse-to-fine | 68.4 | 43.9 | 37.3 | 73.9 |

2). The experiment contains two parts: using the coarse tracking only, and using the entire tracking pipeline. We report the tracking performance in terms of precision, recall, MOTA, and MOTP. MOTA measures the multi-object tracking accuracy, and is a widely metric to evaluate multi-object tracking performance. MOTP measures the multi-object tracking precision, reflecting the average similarity between all true positives and their corresponding ground-truths. The most recent results on MOT2015 are available online (accessed in Sept. 2016)[1], and the state of the art performance is about 76.6% MOTP reported by NOMTwSDP. The results in Table 3 show that the proposed coarse-to-fine tracking works reasonable well (73.9% MOTP) to support the succeeding action cuboid extraction and recognition stages of our multi-action recognition approach.

## 6 Conclusion

We investigate the multiple actions recognition problem in crowd scenarios, which is a problem lacking deep exploration. We use a divide and conquer strategy to handle this problem by proposing a coarse-to-fine multi-object tracking algorithm. Fast tracking using Kalman filter provides action trajectories for most objects with smoothing movement, and fine tracking with sparse optical flow refines the object trajectories with abrupt appearance changes. Such a coarse-to-fine tracking method allows us to obtain individual action tracks with balanced accuracy and speed. Then, traditional action recognition algorithms designed for single action recognition are applied on each of the action tracks in a sliding cuboid mode. Experimental results on a multi-action in crowd database we collected, and the public domain CMU ad MOT15 databases show the proposed approach is effective for actions recognition in crowd video surveillance scenarios.

Accurate spatial-temporal localization of actions still plays a very important role in action recognition. In our future work, we will investigate more robust object detection and tracking approaches for efficient action cuboid localization, and their applications from action recognition, person re-identification to human attribute prediction [48]. Additionally, we will extend the proposed approach towards handling single action recognition tasks in a flexible way.

**Acknowledgement**. This research was partially supported by 973 Program (grant No. 2015CB351802), and Natural Science Foundation of China (grant No. 61672496). The authors would like to thank Xiaoyan Li for her proofreading of

[1] `https://motchallenge.net/results/2D_MOT_2015/`

this paper. H. Han gratefully acknowledges the support of NVIDIA Corporation with the donation of the Titan X GPU used for his research.

## References

1. Siva, P., Xiang, T.: Action detection in crowd. In: BMVC. (2010) 1–11
2. Luo, Y., Cheong, L.F., Tran, A.: Actionness-assisted recognition of actions. In: ICCV. (2015) 3244–3252
3. Li, Y., Ye, J., Wang, T., Huang, S.: Augmenting bag-of-words: a robust contextual representation of spatiotemporal interest points for action recognition. The Visual Computer **31** (2015) 1383–1394
4. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS. (2014) 568–576
5. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with R*CNN. In: ICCV. (2015) 1080–1088
6. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS. (2015) 91–99
7. Fu, Z., Han, Y.: Centroid weighted kalman filter for visual object tracking. Measurement **45** (2012) 650–655
8. Alexei A. Efros, Alexander C. Berg, G.M., Malik, J.: Recognizing action at a distance. In: ICCV. (2003) 726–733
9. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T.S.: Action detection in complex scenes with spatial and temporal ambiguities. In: ICCV. (2009) 128–135
10. Jain, M., Jegou, H., Bouthemy, P.: Better exploiting motion for better action recognition. In: CVPR. (2013) 2555–2562
11. Ryoo, M.S., Matthies, L.: First-person activity recognition: What are they doing to me? In: CVPR. (2013) 2730–2737
12. Zhou, S., Shen, W., Zeng, D., Zhang, Z.: Unusual event detection in crowded scenes by trajectory analysis. In: ICASSP. (2015) 1300–1304
13. Zhu, Y., Nayak, N.M., Roy-Chowdhury, A.K.: Context-aware modeling and recognition of activities in video. In: CVPR. (2013) 2491–2498
14. Li, W., Wen, L., Choo Chuah, M., Lyu, S.: Category-blind human action recognition: A practical recognition system. In: ICCV. (2015) 4444–4452
15. Wu, J., Hu, D., Chen, F.: Action recognition by hidden temporal models. The Visual Computer **30** (2014) 1395–1404
16. Hoai, M., Zisserman, A.: Improving human action recognition using score distribution and ranking. In: ACCV. (2014) 3–20
17. Ni, B., Moulin, P., Yang, X., Yan, S.: Motion part regularization: Improving action recognition via trajectory group selection. In: Proc. CVPR. (2015) 3698–3706
18. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision **103** (2013) 60–79
19. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV. (2013) 3551–3558
20. Jegou, H., Perronnin, F., Douze, M., Sánchez, J., Perez, P., Schmid, C.: Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 1704–1716
21. Chen, W., Corso, J.J.: Action detection by implicit intentional motion clustering. In: ICCV. (2015) 3298–3306

22. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Robust tracking-by-detection using a detector confidence particle filter. In: ICCV. (2009) 1515–1522
23. Choi, W.: Near-online multi-target tracking with aggregated local flow descriptor. In: ICCV. (2015) 3029–3037
24. Chari, V., Lacoste-Julien, S., Laptev, I., Sivic, J.: On pairwise costs for network flow multi-object tracking. In: CVPR. (2015) 5537–5545
25. Ma, C., Huang, J.B., Yang, X., Yang, M.H.: Hierarchical convolutional features for visual tracking. In: ICCV. (2015) 3074–3082
26. Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M.: Learning spatially regularized correlation filters for visual tracking. In: ICCV. (2015) 4310–4318
27. Tang, M., Feng, J.: Multi-kernel correlation filter for visual tracking. In: ICCV. (2015) 3038–3046
28. Liu, T., Wang, G., Yang, Q.: Real-time part-based visual tracking via adaptive correlation filters. In: CVPR. (2015) 4902–4912
29. Wang, L., Ouyang, W., Wang, X., Lu, H.: Visual tracking with fully convolutional networks. In: ICCV. (2015) 3119–3127
30. Bae, S.H., Yoon, K.J.: Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In: CVPR. (2014) 1218–1225
31. Xing, J., Ai, H., Lao, S.: Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In: CVPR. (2009) 1200–1207
32. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly **2** (1955) 83–97
33. Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial intelligence **17** (1981) 185–203
34. Ken Chatfield, Karen Simonyan, A.V.A.Z.: Return of the devil in the details: Delving deep into convolutional nets. In: BMVC. (2014) 2491–2498
35. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. International Journal of Computer Vision **88** (2010) 303–338
36. Kang, D., Han, H., Jain, A.K., Lee, S.W.: Nighttime face recognition at large standoff: Cross-distance and cross-spectral matching. Pattern Recognition **47** (2014) 3750–3766
37. Klum, S.J., Han, H., Klare, B.F., Jain, A.K.: The facesketchid system: Matching facial composites to mugshots. IEEE Transactions on Information Forensics and Security **9** (2014) 2248–2263
38. Han, H., Shan, S., Chen, X., Lao, S., Gao, W.: Separability oriented preprocessing for illumination-insensitive face recognition. In: ECCV. (2012) 307–320
39. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: ICCV. (2007) 1–8
40. Hubel, D., Wiesel, T.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. **160** (1962) 106–154
41. Liu, C., Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans. Image Process. **11** (2002) 467–476
42. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: ICPR. (2004) 32–36
43. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos in the wild. In: CVPR. (2009) 1996–2003

44. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR. (2009) 2929–2936
45. Tran, D., Sorokin, A.: Human activity recognition with metric learning. In: ECCV, Springer (2008) 548–561
46. Ke, Y., Sukthankar, R., Hebert, M.: Event detection in crowded videos. In: ICCV. (2007) 1–8
47. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: Motchallenge 2015: Towards a benchmark for multi-target tracking. arXiv preprint arXiv:1504.01942 (2015)
48. Han, H., Otto, C., Liu, X., Jain, A.K.: Demographic estimation from face images: Human vs. machine performance. IEEE Transactions on Pattern Analysis and Machine Intelligence **37** (2015) 1148–1161