ARC '16

مؤتمر مؤسسة قطر
السنوي للبحوث

QATAR FOUNDATION
ANNUAL RESEARCH
CONFERENCE

Towards World-class
Research and Innovation

Information Communications Technology Pillar

http://dx.doi.org/10.5339/qfarc.2016.ICTPP3076

Action Recognition in Spectator Crowds

Arif Mahmood, Nasir Rajpoot

Qatar University, QA

Email: arif.mahmood@qu.edu.qa

Action Recognition in Spectator Crowds

During the Football Association competitions held in 2013 in UK, 2,273 people were arrested due to the events of lawlessness and disorder, according to the statistics collected by the UK Home Office [1]. According to a survey on the major soccer stadium disasters around the world, more than 1500 people have died and more than 5000 are injured since 1902 to 2012 [2]. Therefore understanding spectator crowd behaviour is an important problem for public safety management and for the prevention of dangerous activities.

Computer Vision is the platform used by researchers for efficient crowd management research through video cameras. However most of the research efforts primarily show results on protest crowds or casual crowds while the spectator crowds have not been focussed. On the other hand the action recognition research has mostly addressed actions performed by one or two actors while the actions performed by individuals in the dense spectator crowds has not been addressed and is still an unsolved problem.

Action recognition in dense crowds pose very difficult challenges mostly due to the low resolution of subjects and significant variations in the action performance by the same individuals. Also different individuals perform the same action quite differently. Spatial distribution of performers varies with time. Scene contains multiple actions at the same time. Thus compared to the single actor action recognition, noise and outliers are significantly large and action start and stop are not well defined making action recognition very difficult.

In this work we target to recognize the actions performed by individuals in spectator crowds. For this purpose we consider a recently released dataset consisting of spectators in the 26th Winter Universiade held in Italy in 2013 [3]. Data was collected during the last four matches held in the same ice stadium using 5 cameras. Three high resolution cameras focussed on different parts of the spectator crowd with 1280×1024 pixel resolution and 30 fps temporal resolution. Figure 1 shows an example spectator crowd dataset image.

Cite this article as: Mahmood A, Rajpoot N. (2016). Action Recognition in Spectator Crowds. Qatar Foundation Annual Research Conference Proceedings 2016: ICTPP3076 http://dx.doi.org/10.5339/qfarc.2016.ICTPP3076.

For action recognition in the spectator crowds, we purpose to compute dense trajectories in the crowd videos by using optical flow [4]. Trajectories are initiated on a dense grid and the starting points satisfy a quality measure based on KLT feature tracker (Figure 2). Trajectories exhibiting motion lower than a minimum threshold are discarded. Along each trajectory shape and texture is encoded using Histograms of Oriented Gradients (HOG) features [5] and motion is encoded using Histogram of Flow (HOF) features [6]. The resulting feature vectors are grouped using the person bounding boxes provided in the dataset (Figure 4). Note that person detectors which are especially designed for detection and segmentation of persons in dense crowds can also be used for this purpose [7].

All trajectories corresponding to a particular person are considered to encode the actions performed by that person. These trajectories are divided into overlapped temporal windows of width 30 frames (or 1.00 second time). Two consecutive windows has an overlap of 66%. Each person-time window is encoded using bag-of-words technique as explained below.

The S-HOCK dataset contains 15 videos of spectator crowds. For the purpose of training we use 10 videos and the remaining 5 videos are used for testing. From the training videos 100,000 trajectories are randomly sampled and grouped to 64 clusters using k-means algorithm. Each cluster center is considered as an item in the code-book. Each trajectory in a person-time group of trajectories is encoded using this code-book. This encoding is performed in the training as well as the test videos using bag-of-words approach. The code-book is considered as a part of the training process and saved.

For the purpose of bag-of-words encoding, distance of each trajectory in the person-time trajectory group is measured from all items in the code-book. Here we follow two approaches. In the first approach, only one vote is casted at the index corresponding to the best matching code-book item. In the second approach, 5 votes are casted corresponding to the 5 best matching code-book items. These votes are given weights inversely proportional to the distance of trajectory from each of the five best matching code-book items.

In our experiments we observe better action recognition performance of the multi-voting strategy compared to the single weight scheme. It is because more information is captures in the multi-voting strategy. In the SHOCH dataset, each person is manually labelled as performing one of the 23 actions, including the ‘other’ action which covers all actions not included in the first 22 categories (Figure 3). Each person-time group of trajectories is given an action label from the dataset. Once this group is encoded using code-book, it becomes a single vector histogram. Each of these vectors is given the same action label depending upon the label assigned to the corresponding person-time trajectory group.

The labelled vectors obtained from the training dataset are used to train both linear and kernel SVM using one verses all strategy. The labels of the vectors in the test data are used as ground truth and the learned SVM are used to predict the label of each test vector independently. The predicted labels are then compared with the ground truth labels to establish action recognition accuracy. We observe an accuracy increase of 3% to 4% when SVM with Gaussian RBF was used. Results are shown in Table 1 and precision recall curves are shown in Figures 5 & 6.

In our experiments we observe that applauding and shaking flag actions have obtained more accuracy compared with other actions in the dataset (Table 1). It is mainly because of the fact that these actions have higher frequency and consist of significant discriminative motion. While other actions have low frequency of occurrence and also in some actions the motion is not discriminative. For example in using device action, when someone in the crowd use a mobile phone or a camera, the motion based detection is not very efficient.

References

[1] Home Office and The Rt Hon Mike Penning MP, “Football-related arrests and banning orders, season 2013 to 2014”, published 11 September 2014.

[2] Associated Press, “Major Soccer Stadium Disasters”, The Wall Street Journal (World), published 1 February 2012.

[3] Conigliaro, Davide, et al. “The SHock Dataset: Analyzing Crowds at the Stadium.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[4] Wang, Heng, et al. “Action recognition by dense trajectories.” Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[5] Dalal, Navneet, and Bill Triggs. “Histograms of oriented gradients for human detection.” Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

[6] Dalal, Navneet, Bill Triggs, and Cordelia Schmid. “Human detection using oriented histograms of flow and appearance.” Computer Vision–ECCV 2006. Springer Berlin Heidelberg, 2006. 428–441.

[7] Idrees, Haroon, Khurram Soomro, and Mubarak Shah. “Detecting Humans in Dense Crowds using Locally-Consistent Scale Prior and Global Occlusion Reasoning.” IEEE TPAMI 2015.