Murat Ambarkutuk {murata@vt.edu}, 03/17/2016

1 Recognizing Realistic Actions from Videos "in the Wild"

In their paper, the authors propose a method of action recognition in which the human actions are recognized from unconstrained videos sequences. The contributions of the paper relax the assumptions declared in earlier literature, in particular illumination, camera motion, view point restrictions and clutter, which makes the algorithms more applicable to realistic videos than the earlier proposed methods. The proposed approach employes two different features capturing different statistics of the video sequence of interest.

The first feature set, referred as "Local Motion Features" in the paper, captures local spatiotemporal interest points. The interest points then are quantized and vectorized into statistical distributions which represents the video sequences. The authors point out that humans can recognize the actions from sequences still images depicting human poses. The second feature set (Local Static Features) imitates the human cognitive system in a way that the context can be recovered from sequences of poses. The authors assert that this feature set accounts for "the wild" nature of realistic videos by compensating unintended camera motion such as shake. However these two feature sets can be regarded as two different solutions for two different problems, the authors signifies that it is essential to use both information sources to deal with the drawbacks of both feature representations while making use of their advantages. Two cases are underlined the importance of these complementary features in the paper. The first case was that it is really difficult for the motion features to distinguish biking from horse-riding, because both sequences will show the same statistics. On the other hand, it is difficult for the static features to distinguish jogging from running, because the poses are in the both frames are about the same. The motion features are pruned to eleminate those belong the background while PageRank algorithm is used to select important the static features. The visual-vocabulary is then learned with an information-theoretic algorithm. The learning process contains two phase clustering algorithm, k-means and KL-divergence like divisive algorithm. Then Adaboost algorithm is used to combine two different information sources by boosting for action recognition.

2 Results and Conclusions

The algorithm is tested with two different dataset; KTH, which is a structered action recognition dataset, while the other is a collection of YouTube videos and in-house built video sequences. The second dataset is extensive in size and less structured.

The KTH dataset is utilized to evaluate two aspects of the proposed method. The first aspect assessed is to what extent the combination of the feature sets improves the overall performance. The authors reported that combining two features sets resulted in more than 4 and 9 percent improvement for motion and static feature set, respectively. The second aspect of the method evaluated is how well the vocabularies are described under different number of words. It is denoted that the proposed method is more accurate than k-means alone under different number of words, in particular it shows better performance with fewer words.

The second experiment conducted on in-house built dataset to assess how pruning processs effects the overall approach. Along with these aspects, the hybrid features are re-evaluated with this dataset for validition. The pruning of motion features show about 8% improvement in overall recognition accuracy, while it is worse or about the same in the swinging and the diving categories. Furthermore, the feature mining improves the recognition task about 5%-25% depending on the

categories. While the feature mining is superior than the baseline in horse riding category about the 25%, in the v_spiking, swinging and diving categories it shows relatively poor performance. It is surprising for me to see that, the diving category is the most accurately recognized category while both feature mining and pruning tends to have problems (to some extent) with it.