https://doi.org/10.2352/ISSN.2470-1173.2018.2.VIPC-254

Prediction system for activity recognition with compressed video[1]

Chengzhang Zhong, Amy R. Reibman - Purdue University, West Lafayette, IN, U.S.A

Abstract

Executing video analytics tasks using a large camera network is a challenging problem in the field of video processing. Video compression is a necessary step to reduce video data size before transmission. However, the performance of video analytics tasks generally degrade as video quality drops. This paper considers how to find the optimal point between video compression and performance for the video analytics task of activity recognition. We propose a system that predicts the success or failure of a video analytics task under different compression parameters without executing the task. The system is designed to automatically select the best compression rate for each video to maintain an acceptable detection accuracy. Our experiments indicate that such a system has the potential to improve overall performance across a variety of different activity sets selected from the UCF-101 dataset [1].

1. Introduction

In recent years, millions of devices have been connected to the Internet. The management of this video content becomes a problem due to its size. Video analytics is one of the approaches to discover the hidden information in videos [2] and can be used to detect objects and activities which can be used to make decisions. Video analytics can contribute to applications such as surveillance systems, data storage and management systems. The combination of executing video analytics tasks in a large camera network is a promising as well as challenging task.

Systems based on large amounts of video such as video surveillance systems normally include a bandwidth constraint. This constraint limits the video quality during transmission. For video analytics tasks, the detection accuracy usually drops as video quality degrades. The goal of our proposed system is to efficiently identify, at the edge of the network, the optimal amount of compression to minimize bandwidth during transmission while maintaining a reasonable detection accuracy.

Several video analytics tasks have been explored in the context of compression and its impact. In [3, 4], there is an experimental-based discussion on how to find the trade-off point for face detection and face tracking tasks from compression. In [5], they discussed combined algorithms to focus compression resources on certain interesting elements for the task of object tracking. In [6], the focus was on modeling object detection's performance with the change of Quantization Parameter (QP) in compression. Through a fitting method, a relation between QP, false positive pixels (FP) and false negative pixels (FN) can be built. But the prediction of FP and FN from different QPs is not a model that can be applied to other video analytics tasks. In [7], a system was designed to use texture descriptors to predict detection accuracy of a pedestrian detection task. This system has demonstrated its ability to save data rate and computational resources. However, none of these have considered the impact of compression on activity recognition.

In this paper, we examine the impact of compression on detection accuracy in activity recognition. We explore this using different sets of activities, and show that each activity is affected by compression differently and that the impact of compression depends on the "neighboring" activities from which this activity is to be distinguished. Moreover, we propose a video analytics system corresponding to the task of activity recognition using compressed videos. We use feature descriptors to predict the success or failure of the activity recognition task under different QP values. With this prediction result, the system then selects an optimal compression rate for each input video. This can enable an acceptable detection accuracy and video data bitrate to be achieved.

The paper is organized as following: Section 2 provides the overview of activity recognition. Section 3 explores the impact of compression on activity recognition. Section 4 describes the design details of the compression-aware video analytics system. Section 5 demonstrates the system can perform better than one that uses the same QP to compress all videos and section 6 concludes the paper.

2. Activity recognition

The goal of activity recognition is to recognize human activity from an input video. In other words, the idea is to classify the input video into its activity category [8]. In [9], a Harris point operator was designed to detect spatial-temporal interest point (STIP). Through extracting space-time features and using a Support Vector Machine (SVM), the experiments in [10] were able to achieve a state-of-the-art result on KTH dataset [11]. As the pipeline bag-of-features became popular, an evaluation of different feature detectors, descriptors under common experimental setting was presented in [12]. The result indicates the Histogram of Oriented Gradients (HOG) [13] and the Histogram of Oriented Flow (HOF) [14] are effective features for activity recognition.

In this paper, we use the **Improved Dense Trajectory (IDT)** [15] for the activity recognition task using the implementation from [16]. For each input video, **IDT** applies dense sampling to select interest points at different image scales for each frame. Between consecutive frames, a dense optical flow field is used to estimate the motion interest points. The position of each interest point is recorded and updated to construct their trajectories. Along with the shape of each trajectory, HOG, HOF and Motion Boundary Histogram (MBH) [14] features are extracted from 3D volumes. A technique of warping between consecutive frames is

[1] This work was supported by the Cisco University Research Program Fund CG-#594368 through the Silicon Valley Community Foundation

also applied to eliminate the effect of camera motion. After feature extraction, each video's feature descriptor is encoded through Fisher Vector [17]. In the final step, a SVM will process these encoded features for training or testing.

Typically, to evaluate the performance of a method, there are some public available datasets for activity recognition: The KTH dataset [11] is one of the easy and common dataset. The Hollywood2 dataset [18] includes video clips from movies of actions and scenes. Other dataset such as the UCF sports [19], UCF 50 [20] and UCF101 [1] datasets are built using realistic action videos. In this paper, we examine performance using several subsets of the UCF101 [1] database. To explore the impact of the set of activities, we use different subsets of 10 activities chosen among the entire collection of 101 activities. We create five different activity sets which are listed in Table 1. There are small variations between the activities chosen for **Sets A**, **B**, **C** and **D**, while in **Set E**, most of the activities are completely different from the other sets. For example, **Sets A** and **B** differ only in the activity **"Juggling Balls"** and **"Rowing"**, while there are three different activities in **Set C** relative to **Set A**.

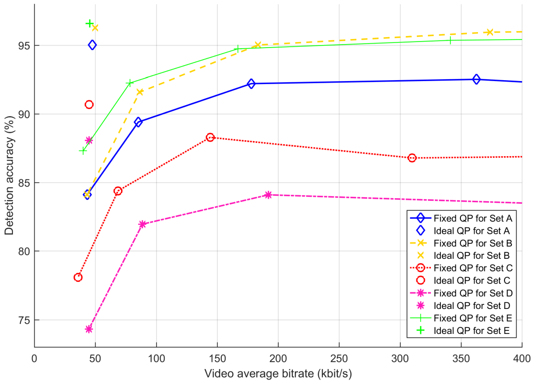


Figure 1: *Activity Sets A – E: ideal QP points and fixed QP curves*

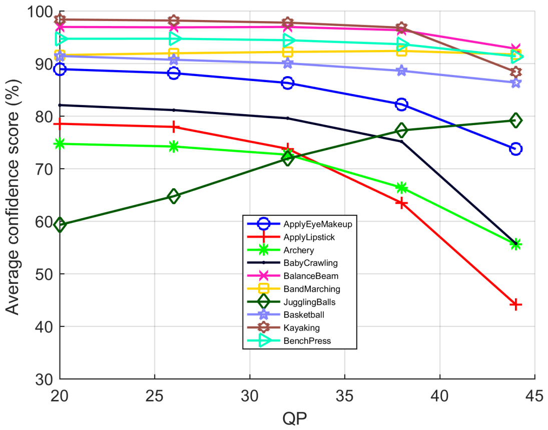


Figure 2: *Set A: per-activity confidence score versus QP*

Table 1: *Activity names of five different sets*

Activity Set A	***Activity Set B***	***Activity Set C***	***Activity Set D***	***Activity Set E***
ApplyEyeMakeup	ApplyEyeMakeup	ApplyEyeMakeup	Archery	ApplyEyeMakeup
ApplyLipstick	ApplyLipstick	ApplyLipstick	BabyCrawling	BreastStroke
Archery	Archery	Archery	BandMarching	Fencing
BabyCrawling	BabyCrawling	BabyCrawling	HorseRace	Haircut
BalanceBeam	BalanceBeam	BalanceBeam	JugglingBalls	IceDancing
BandMarching	BandMarching	BandMarching	MoppingFloor	MilitaryParade
JugglingBalls	Rowing	JugglingBalls	PlayingSitar	PlayingDhol
Basketball	Basketball	PlayingCello	Punch	SalsaSpin
Kayaking	Kayaking	PlayingSitar	Rowing	TaiChi
BenchPress	BenchPress	Rowing	YoYo	WalkingWithDog

3. Impact of compression

To investigate how video compression influences activity recognition, we encode each video using Mencoder [21] into five different compressed versions in H.264 format, each using a constant QP (Quantization Parameter) from among the list {20,26,32,38,44}. A large QP value generates a low bitrate, which results in low video quality. In our designed method, we train only one SVM for each activity class. This SVM depends on each specific set of activities, and is used for all QP values.

Figure 1 demonstrates the detection accuracy for each set, when a fixed QP is used to compress all videos (where the points for QP=20 are truncated because the average bit-rate is higher than 400 kbit/s). As it is indicated in Figure 1, the general trend of the fixed QP curves shows decreasing performance as the bitrate drops, especially below 100 kbit/s. Performance differs significantly across the different test sets. In addition, there exist a few surprises. For instance, on the fixed QP curve of **Set C**, the performance degrades as the bitrate increases in the middle range of the curve.

To understand more about these observations, we analyze the impact of video compression on each activity class in **Set A**. The impact of compression for each activity is evaluated as the average of all test videos' confidence score from the respective SVM. The results are demonstrated in Figure 2 and Figure 3. Figure 2 represents detection results from each individual activity class in **Activity set A**. It is obvious as the video quality degrades (QP increases), most actions' detection confidence scores decrease, but each to a different degree. In Figure 2, the confidence score for the activity **"ApplyLipstick"** drops 34% as the QP increases from 20 to 44. However, the confidence score for activity **"BalanceBeam"** drops only 4%. Moreover, the activity **"JugglingBalls"** increases its confidence score as the compression increases from QP value 20 to QP value 44. This implies that compression actually makes this activity class easier to identify, and accounts for the decrease in accuracy across the entire set shown in Figure 1 as the QP increases from 38 to 44. Therefore, from Figure 2, we see that for the task of activity recognition, the impact of compression depends heavily on the specific activity class.

We also observe that different combinations of activity

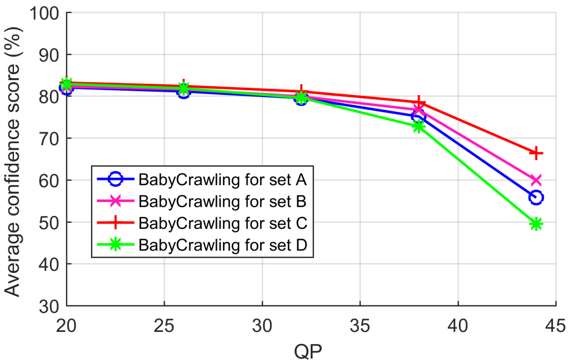


Figure 3: *"BabyCrawling" confidence score versus QP*

classes influences detection accuracy. **Set A**, **Set B**, **Set C** and **Set D** all include the activity class **"BabyCrawling"**. As shown in Figure 3, all of them start with the same confidence score around 81%, but as the QP increases to 44, the confidence score of class **"BabyCrawling"** in **Set A**, **Set B**, **Set C** and **Set D** decreases by 26.3%, 22.3 %, 16.7% and 33.3%, respectively. This indicates that the impact of compression on the detection performance of each activity depends on the set of other activities.

As these results show, when considering the impact of compression on activity recognition, it is important to consider the impact both on each individual activity and due to different collections of activities. Therefore, in this paper, we propose a system that predicts the optimal amount of compression for each individual video.

Consider a system that could compress each individual video i using an ideal quantizer, QP_i^*, that corresponds to the largest QP in our list that produces the correct detection result for that video. While such a selection may not be possible in practice, performance of such a system can demonstrate whether overall performance could be improved if the amount of compression could be optimally chosen for each input video.

To demonstrate the power of such a system, we define the concept of ideal QP point. The ideal QP point represents the result where all the test videos are compressed to the lowest quality to be detectable by the "Improved Dense Trajectory" (IDT) activity recognition algorithm. The ideal points for each set are also shown in Figure 1. Compared to the fixed QP curves, the ideal points demonstrate promising performance both in bitrate saving and accuracy improvement. However, these ideal points are obtainable only with perfect knowledge. Therefore they provide the upper bounds on the performance of our system.

4. Prediction system

In this section, we present our prediction system whose goal is to predict each input video's performance under different compression QP values and select the optimal QP value for each video. This prediction would operate at or near the camera location, and is designed to be light-weight processing with low computational requirements. The proposed system is illustrated in Figure 4. The main components are: Feature extraction, Hierarchical K-means, Random Forest and Compression rate selection. The system starts with feature extraction from all compressed versions of the input video, and then applies the visual word assignment pipeline [22] to assign words to each descriptor. The resulting histogram represents the video. After that, the histogram is input to the trained Random Forest to receive a classification result whether the detection performance is **"success"** or **"failure"** for the given QP value. The final step is to collect the classification results from the previous step and select the optimal QP. The following sections will describe each component in detail.

Feature Extraction

Texture features normally include representative information about the video. In order to find an appropriate feature for the prediction system, we evaluated four different types of features in this paper. According to [12], densely sampled features have the best performance on complex datasets. Therefore, all the features evaluated in this paper are densely sampled. We selected HOG (Histogram of oriented gradients) [13, 22], HOF (Histogram of oriented flow) [14, 22], MBH (Motion Boundary Histograms) [14, 22] and SIFT (Scale-invariant feature transform) [23, 24] to test our prediction system.

Visual word assignment

Densely sampled features extracted from different videos normally have a different dimension of descriptors due to the video length. In our system, we choose to use hierarchical k-means [22] to assign each video with an equal length histogram of visual words.

Random Forest

A Random Forest [25] includes a collection of decision trees, where the growth of trees and the split of nodes both depend on random selection. An input vector proceeds through each tree to receive a decision vote. After collecting the votes across all trees, the forest selects the class that receives the most votes as the final decision. In our prediction system, for a given set of activities, we design five Random Forests, one for each QP value considered. Each Random Forest predicts success or failure of the activity recognition task for a given input video compressed using that QP.

To train each Random Forest for this classification problem, we require feature inputs and the correspond labels. The input feature for one video is the histogram of visual words from the last step. For labels, we use the confidence score of the SVM predictor in the activity recognition task. The score indicates the probability that a video belongs to its ground truth activity class. In our system, if the score is higher than 0.5, we denote it as **"success"**. Otherwise, we denote it as **"failure"**.

From the perspective of training our system, we need to train both the activity recognition **IDT** algorithm and our prediction system. We split the 25 groups of videos inside each activity class into two parts. Group 01 to 12 is used for testing the **IDT** algorithm and 13 to 25 for training the **IDT** algorithm. Furthermore, to avoid training the Random Forests on the same data used to train the **IDT**, we randomly split groups 01 to 12 into two parts equally: one part for training the Random Forests, another part for testing them. The prediction result for each video is a 1x5 vector indicating **"failure"** or **"success"**, where each element in the vector corresponds to one QP value.

Compression rate selection

After each Random Forest predicts whether a correct decision will be made at each QP considered, we select the estimated $\hat{QP}_i$ to be the largest QP that yields a **"success"** prediction.

Examining the fixed QP curves in Figure 1, we notice that

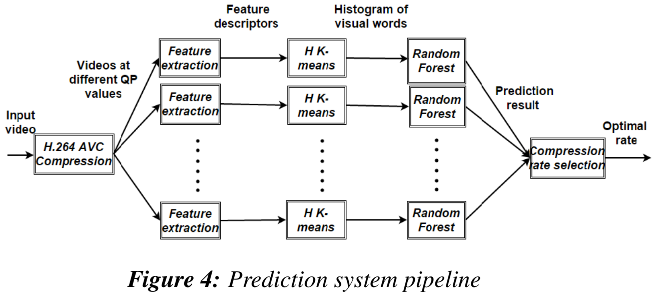


Figure 4: *Prediction system pipeline*

Table 2: *Random Forest training and testing samples for Set A and E*

	success	failure
Number of Set A Training samples	1365	230
Number of Set A Testing samples	1437	168
Number of Set E Training samples	1464	136
Number of Set E Testing samples	1499	116

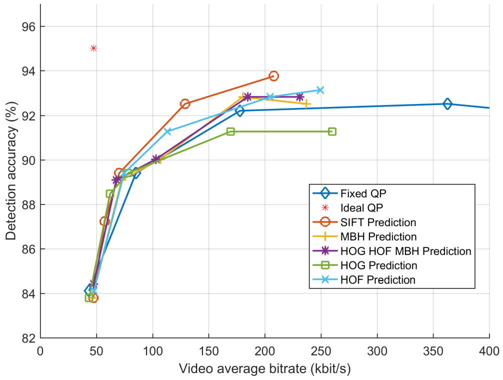


Figure 5: *Prediction result from Set A*

the detection accuracy drops gradually at high bitrate and sharply at low bitrate as the bitrate decreases. Therefore, it is usually good to conservatively trade a small amount of bitrate for a relatively large detection accuracy increment. Therefore, in some extreme cases, when all QPs lead to a **"failure"** prediction, we conservatively compress these videos using QP 20, which corresponds to the highest video quality and bitrate.

5. Experimental Result

In this section, we construct an evaluation method whose goal is to test the performance of our prediction system. The prediction system's pipeline has already been discussed in section 4. However, for the activities and QPs that we consider, the overall number of samples are quite unbalanced between **"success"** and **"failure"** for each QP, as shown in Table 2. The **"failure"** cases comprise only a small proportion of the total samples. This is a crucial factor that will limit our prediction accuracy [26]. To reduce the impact of this imbalance, we pre-assigned class weights to the Random Forest during training to increase the importance of the minority class. As the assigned class weight varies, we plot each feature's prediction result as a curve.

These results are shown in Figures 5 and 6, for the different features mentioned in section 4. The figures examine the performance of an entire activity recognition system that incorporates our prediction results, for activities **Set A** and **Set E** respectively. Recall that these two sets contain only the activity **"Apply Eye Makeup"** in common. For **Set A**, our system performs better than the system with a fixed QP for almost all of the features we tried, in the range from 50 to 150 kbit/s. The best performance is for SIFT in the middle range of bitrates. For **Set E**, our prediction system performs better than a fixed-QP system only when the bitrate is less than 100 kbit/s. All features have nearly equivalent performance for bit-rates around 70 kbit/s. For both sets of activities, however, the accuracy of the system is still significantly lower than the ideal point.

To interpret the result in Figure 5 and 6, we examine confusion matrices of the ideal QP value (horizontal axis) and predicted QP value (vertical axis) in Figures 7 and 8, respectively. Both matrices reflect the prediction point of highest accuracy. For **Set A**, this is the right-most point on the SIFT curve. For **Set E**, this is the right-most point on the HOG curve. The column labeled **"No activity detected"** indicates when the original activity recognition system is unable to identify the activity, regardless of the amount of compression. The row labeled **"Task failure"** indicates when our prediction system predicts that there is no QP for which an activity will be detected. Note that for those videos for which our system predicts a **"Task failure"**, our system conservatively compresses these videos with QP=20, with the goal to increase the chance that the activity is correctly detected.

A video whose ideal QP is larger than the predicted QP will require our system to spend additional bitrate without increasing the detection accuracy. This is also true for those videos for which our system predicts a **"Task failure"**. A video whose ideal QP is smaller than the predicted QP may save the system bitrate but at the cost of a decrease in detection accuracy. Therefore, by examining these confusion matrices we can gain insight into our system performance.

Several observations are apparent from Figures 7 and 8. First, examining the "ideal QP" for both sets, we see that a majority of videos can be compressed heavily with QP=44 while still being able to identify the activity. When our system correctly predicts a QP=44, these videos are optimally compressed for the

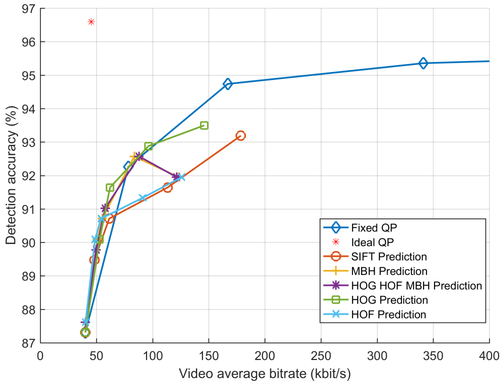
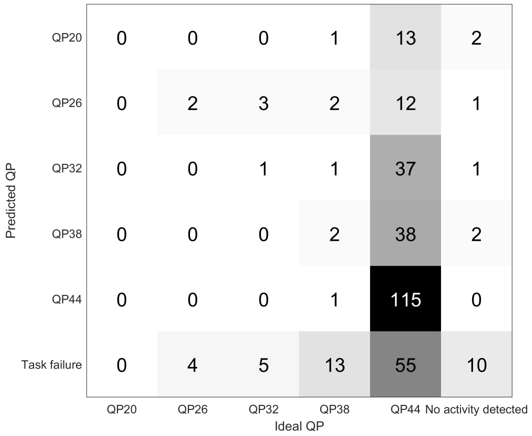
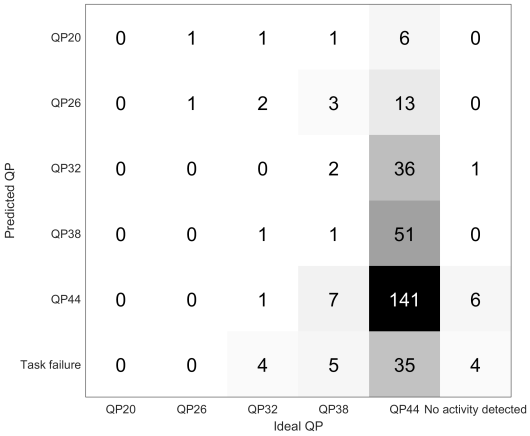


Figure 6: *Prediction result from Set E*



	QP20	QP26	QP32	QP38	QP44	No activity detected
QP20	0	0	0	1	13	2
QP26	0	2	3	2	12	1
QP32	0	0	1	1	37	1
QP38	0	0	0	2	38	2
QP44	0	0	0	1	115	0
Task failure	0	4	5	13	55	10

Figure 7: *Confusion matrix for SIFT from Set A*



Predicted QP	QP20	QP26	QP32	QP38	QP44	No activity detected
QP20	0	1	1	1	6	0
QP26	0	1	2	3	13	0
QP32	0	0	0	2	36	1
QP38	0	0	1	1	51	0
QP44	0	0	1	7	141	6
Task failure	0	0	4	5	35	4

Figure 8: *Confusion matrix for HOG from Set E*

activity recognition. However, for these videos, choosing a less aggressive QP will increase the bitrate of our system without improving detection accuracy. In particular, for **Set A** in Figure 7, we see that while 270 videos have an ideal $QP^* = 44$, 100 of these videos are predicted by our system to require a finer QP, unnecessarily increasing the bitrate. Moreover, the 55 videos with ideal $QP^* = 44$ for which our system predicts a task failure will also require increased bitrate based on our conservative compression strategy. Similarly, for **Set E** in Figure 8, 282 videos have an ideal $QP^* = 44$, of which 106 are conservatively compressed; moreover, 35 videos are conservatively compressed due to **"Task failure"**. Combined, this has the effect of moving the "ideal point" shown in Figures 5 and 6 to the right, without reducing the detection accuracy.

However, we also can observe from Figures 7 and 8 that there are a number of videos that require a less-aggressive QP than 44 for correct activity detection, for which our system predicts too aggressive a QP, relative to the ideal QP^*. This significantly reduces detection accuracy while saving only a small amount of bit-rate. All aggressively-predicted videos appear in the lower triangular region below the diagonal in the matrices of Figures 7 and 8. Comparing these matrices, we see that the percentage of aggressively predicted videos for **Set A** and **Set E** are 0.31 % and 2.78 %. The greater fraction of these videos in **Set E** relative to **Set A** explains why our system performs better for the latter than the former.

As pointed out above, when our system predicts a **"Task failure"**, the associated video is conservatively compressed, increasing the bitrate without improving the detection accuracy. The percentage of **"Task failure"** videos for **Set A** and **Set E** are 27.10 % and 14.86 %. We believe **"Task failures"** are so frequent because our prediction system is limited by the imbalanced training data, as discussed earlier. Improvements to the training data could improve our overall system performance.

We also compare the performance of each feature in our prediction system. For example, in Figure 5, SIFT feature has a better detection accuracy compared with other features over all bitrates. Therefore, SIFT has the best performance for **Set A**. However, after checking the performance of each feature in all sets, we notice that the HOF feature has the best performance in three out of five of these activity sets. Thus while more exploration is needed, the HOF may be the most reliable feature for our prediction system.

6. Conclusion

In this paper, we proposed a system to predict each video's optimal compression rate for the task of activity recognition. The goal was a system with lightweight processing at the edge of the network, located near the camera, that could achieve the lowest bitrate possible without sacrificing detection accuracy. Toward this end, we explored the effect of compression on the performance of activity recognition using different sets of activities, and demonstrated that significantly different trends are present, depending on the composition of the set of activities. To evaluate the performance of our system, we defined the concept of an ideal QP point, which indicates an upper bound on our performance. The ideal QP point indicates a great deal of promise that significant gains might be possible in both detection accuracy and bitrate. Through our experiments, we were able to generate an acceptable prediction result for some of the combinations of activities. For other combinations, we analyzed the potential factors which limit the performance of our system. Because the results of our system lie significantly below the ideal QP point, there still exists potential space to explore in this field.

References

[1] Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah."UCF101: A dataset of 101 human actions classes from videos in the wild." arXiv preprint arXiv:1212.0402 (2012).

[2] Choudhary, Ayesha, and Santanu Chaudhury. "Video analytics revisited." IET Computer Vision 10.4 (2016): 237-247

[3] Korshunov, Pavel, and Wei Tsang Ooi. "Video quality for face detection, recognition, and tracking." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 7.3 (2011): 14.

[4] Korshunov, Pavel, and Wei Tsang Ooi. "Critical video quality for distributed automated video surveillance." Proceedings of the 13th annual ACM international conference on Multimedia. ACM, 2005.

[5] Soyak, Eren, Sotirios A. Tsaftaris, and Aggelos K. Katsaggelos. "Low-complexity tracking-aware H. 264 video compression for transportation surveillance." IEEE Transactions on Circuits and Systems for Video Technology 21.10 (2011): 1378-1389

[6] Kong, Lingchao, Rui Dai, and Yuchi Zhang. "A new quality model for object detection using compressed videos." Image Processing (ICIP), 2016 IEEE International Conference on. IEEE, 2016.

[7] Tahboub, Khalid, Amy R. Reibman, and Edward J. Delp. "Accuracy prediction for pedestrian detection." 2017 ICIP, September.

[8] Aggarwal, Jake K., and Michael S. Ryoo. "Human activity analysis: A review."ACM Computing Surveys (CSUR) 43.3 (2011): 16.

[9] Laptev, Ivan, and Tony Lindeberg. "Space-time interest points." 9th International Conference on Computer Vision, Nice, France. IEEE conference proceedings, 2003.

[10] Laptev, Ivan, et al. "Learning realistic human actions from movies." Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008.

[11] Schuldt, Christian, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: A local SVM approach." Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Vol. 3. IEEE, 2004.

[12] Wang, Heng, et al. "Evaluation of local spatio-temporal features

for action recognition." BMVC 2009-British Machine Vision Conference. BMVA Press, 2009.

[13] Dalal N., and B. Triggs, "Histograms of oriented gradients for human detection." IEEE International Conference on Computer Vision and Pattern Recognition, 2005.

[14] Dalal, Navneet, Bill Triggs, and Cordelia Schmid. "Human detection using oriented histograms of flow and appearance." European conference on computer vision. Springer, Berlin, Heidelberg, 2006

[15] Wang, Heng, and Cordelia Schmid. "Action recognition with improved trajectories." Proceedings of the IEEE International Conference on Computer Vision. 2013.

[16] "DTF + Fisher Vector code", URL:http://www.boyang.net/2014/04/30/fisher-vector-in-action-recognition

[17] Perronnin, Florent, Jorge Sanchez, and Thomas Mensink. "Improving the fisher kernel for large-scale image classification." Computer Vision:ECCV 2010 (2010): 143-156.

[18] Marszalek, Marcin, Ivan Laptev, and Cordelia Schmid. "Actions in context." IEEE International Conference on Computer Vision and Pattern Recognition, 2009.

[19] Rodriguez, Mikel D., Javed Ahmed, and Mubarak Shah. "Action mach a spatio-temporal maximum average correlation height filter for action recognition." Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008.

[20] Reddy, Kishore K., and Mubarak Shah. "Recognizing 50 human action categories of web videos." Machine Vision and Applications 24.5 (2013): 971-981.

[21] "Mencoder", URL: http://www.mplayerhq.hu/design7/news.html

[22] Uijlings, J., et al. "Video classification with densely extracted HOG/HOF/MBH features: an evaluation of the accuracy/computational efficiency trade-off."International Journal of Multimedia Information Retrieval 4.1 (2015): 33-44.

[23] Uijlings, Jasper RR, Arnold WM Smeulders, and Remko JH Scha. "Real-time visual concept classification." IEEE Transactions on Multimedia 12.7 (2010): 665-681.

[24] Lowe, David G. "Distinctive image features from scale-invariant keypoints."International Journal of Computer Vision 60.2 (2004): 91-110.

[25] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[26] Chen, Chao, Andy Liaw, and Leo Breiman. "Using random forest to learn imbalanced data." University of California, Berkeley 110 (2004)

Author Biography

Chengzhang Zhong is a PhD student working under Professor Amy R. Reibman in the school of Electrical and Computer Engineering at Purdue University. He received his bachlor degree at Purdue University (2015).