Content Based Video Quality Control for Wide-area Video Surveillance Systems

Takeshi Arikuma[1], Ashish Jain[2], Kazuya Koyama[3], Kang Wei Woo[4],
Tsunehisa Kawamata[1] and Keiji Yamada[1]

[1] NEC Asia Pacific Pte. Ltd.
No.1 Maritime Square
#12-10,
Singapore

[2] NEC Technologies India
Plot No. 20 and 21,
Sector-135,
Noida, India

[3] NEC Corporation
7-1, Shiba 5-chome,
Mitato-ku,
Tokyo, Japan

[4] NEC Corporation
Singapore Branch
No.1 Maritime Square
#12-10,
Singapore

takeshi_arikuma@nec.com.sg, ashish.j@nectechnologies.in, k-koyama@ax.jp.nec.com,
kangwei_woo@nec.com.sg, tsunehisa_kawamata@nec.com.sg, keiji_yamada@nec.com.sg

ABSTRACT

In accordance with the increase of the demands on wide-area surveillance to minimize the damage for incidents such as terrorism and riots, the importance of the sharing of video footage among local surveillance hubs is increasing to take action quickly and effectively.

In this paper, we propose a system which can reduce the network usage between local surveillance hubs by changing bitrates depending on contents of the video feeds. The system was designed to support various camera configurations available in a real surveillance setup by using a middleware called ASCOT, on which surveillance systems can be developed by assembling different types of video analyses.

As an evaluation, we developed crowd congestion detection, face detection and motion detection on top of the middleware and evaluated them with two footages taken in a real important premise, one from entrance and another one from corridor. Crowd congestion detection reduced 41% of network usage for the entrance video footage, and face detection reduced 46% of network usage for the corridor video footage where motion detection reduced 6% and 27% respectively. Furthermore, we confirmed that the system can apply these video analyses functions and flexibly combine these analyses for video streams.

KEYWORDS

Large area surveillance, image processing, surveillance middleware, bandwidth management

1 INTRODUCTION

The increase of wide-area incidents such as terrorism and riots is one of important social problems. To prevent and minimize the damage for these incidents, the expectations for changing the traditional local area surveillance, such as premises or districts level, to wide area surveillance, such as whole city or state level, is increasing. This kind of wide area surveillance includes information sharing between existing local surveillance hubs (local hubs) and new "wide-area surveillance hubs" (wide-area hubs) which take care of terror or widespread disaster to make it possible for authorities to react to widespread incidents quickly.

In wide-area hubs, video analysis over multiple camera feeds became one of key feature. They need to monitor thousands of feeds from local hubs and the correlation of activities in multiple places will be important to detect terror or disaster. In other words, complete distributed system is not suitable because processing multiple camera feeds require data transfers between local hubs and result in large data transfers over a mesh network of local hubs. We also need to consider infrastructure limitations and requirement of the surveillance operation in real world.

We modeled the system as shown in Figure 1 to analyze and clarify the requirements and constrains. In the model, the original video feed from cameras will be stored in local hubs and

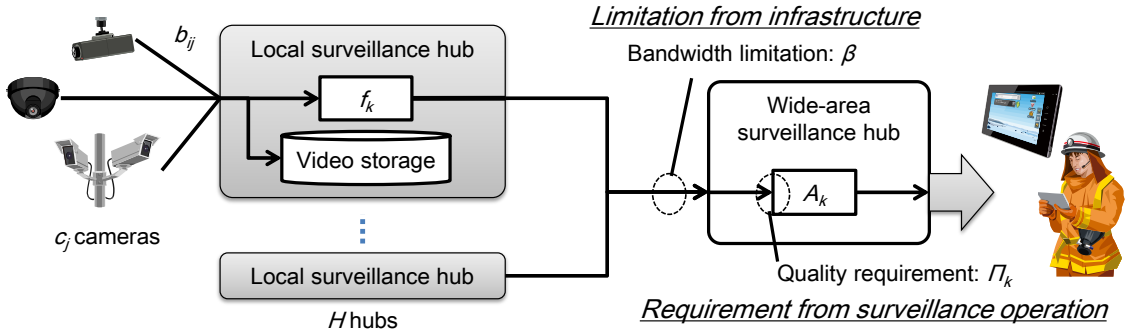

Figure 1 Wide-area surveillance system model

local hub will send video feed after data compression f_k to reduce the data transferred to the wide-area hub. In wide-area hub, video analyses (A_k) are executed over multiple cameras to realize end user use cases. Examples of A_k includes face matching, intrusion detection and crowd behavior analysis.

As for the infrastructure, one of the large constrain is network bandwidth. The typical available lease line is 1Gbps to 10Gbps in developed countries, which is not enough to receive the video feeds from existing local hubs. For example, Singapore's Mass Rapid Transit (SMRT) has about 6,000 cameras in 49 stations for two lines [1]. There will be more than 12,000 cameras in whole metro system in Singapore (106 stations and five lines). The necessary bandwidth to share these video feeds is estimated as 18Gbps (1.5Mbps / camera), which is difficult to send over lease lines.

We can formulate this constrain in the model as following where b_{ij} is the original video bitrate from camera i in the local hub j, f_k is the compression function in local hub, c_j is the number of the camera in the local hub j, H is the number of local hubs, and β is the bandwidth limitation.

$$\sum_{j=1}^{H}\sum_{i=1}^{c_j} f_k\left(b_{ij}\right) \le \beta \qquad (1)$$

Another requirement from surveillance operation is the quality of video feeds. In the wide-area hub, support of video analysis technologies to find target object, person or behavior become important because it should handle thousands of feeds from local hubs in real-time. It is important to use high quality video for these video analyses because video compressions can deteriorate the video quality and results in the drop of video analysis accuracy [2, 3].

We can formulate this requirement as following where A_k is real-time video analysis, and Π_k is the accuracy requirement on the video analysis such as acceptable false alarms.

$$accuracy\left(A_k\left(f_k\left(b_{ij}\right)\right)\right) \ge \Pi_k \qquad (2)$$

The objective of this research is to find the compression function (f_k) which fulfills both (1) and (2). One of the candidates of the solution for this problem is content base quality control. In this approach, f_k compress the video by two steps: labels video segments with importance of the content base labeling algorisms such as motion detection, face detection or congestion detection, and changes the bitrate according to the labels of the segment. Motion detection used in surveillance cameras is one of the examples for this technique.

Though content base quality control is a potent approach, we need to answer two questions to utilize this approach in real surveillance systems: how much we can compress the data in real surveillance setup and how we can handle the situation when most of the video segments are marked important and the inequation (1) might not be fulfilled.

In this paper, we focus on the first question, i.e. how much we can compress the data in real setup with content base quality control.

We propose and evaluate a system which compresses the data for wide range of surveillance camera configurations by

ISBN: 978-1-941968-02-4 ©2014 SDIWC

designing a system so that multiple labeling algorisms can be selected and assembled as compression function to meet the use cases of each camera. We apply crowd congestion detection and face detection which can be used for congested video footages as examples of the preprocessing in local hubs. We also discuss about the situation when most of the video segments are marked important and clarify what is needed as next step.

The rest of the paper is organized as follows. In section 2, we review existing network usage reduction technologies. Section 3 clarifies requirements for the content base quality control system in wide area surveillance applications. Section 4 discusses our approach to fulfill the requirements described in section 3. In section 5, we described the methods for the evaluation and results obtained. Section 6 concludes this paper with a summary of key points and a mention of future works.

2 RELATED WORKS

The studies on network usage reduction using predefined rules or simple video analysis as labeling algorism for compression function has been done. Pillai et.al proposed a storage reduction technique using predefined rules such like reduce the resolution for every two frames [4]. Motion detection is used in surveillance cameras and video management software to reduce the resource usage [2]. Korshunov et.al proposed an approach to use face detection to reduce the network usage which showed in 72-92% reduction with in-lab video [5]. These studies showed that the network usage can be significantly reduced by changing resolution based on labeling algorisms for some specific use case or camera configuration.

However, these technologies have strong constrains on use case and camera configurations. For example, motion detection can compress sparse video. Face detection can be applied for cameras which installed with low depression angle and capture the person features large enough.

To cover the variety of camera configurations in real surveillance systems, it is important to find technologies which can applicable for this kind of camera configurations. For example, crowd analysis, which has been intensively researched last decade [6, 7], is one of the candidates of labeling algorisms that can be applicable to cameras in crowded areas.

3 REQUIREMENTS ON CONTETN BASE QUALLITY CONTROL

To fulfill constrains (1) and (2) discussed in introduction, there are two requirements on content base quality control technology, i.e., support of crowded and panoramic view cameras, and the flexibility to support variety of camera configurations and use cases.

3.1 Support of Crowded Videos

It is important to support crowded videos because not negligible numbers of surveillance cameras are installed in crowded places such as lobby, corridor and entrances. An example of this kind of video is shown in Figure 2. These cameras are used to:

- prevent accidents caused by overcrowding such as fall down on top of one another
- limit the number of people for safe evacuation in case of fire

Since these use cases are common in surveillance systems, it is important to clarify what kind of labeling algorism can be applicable for these crowded videos.

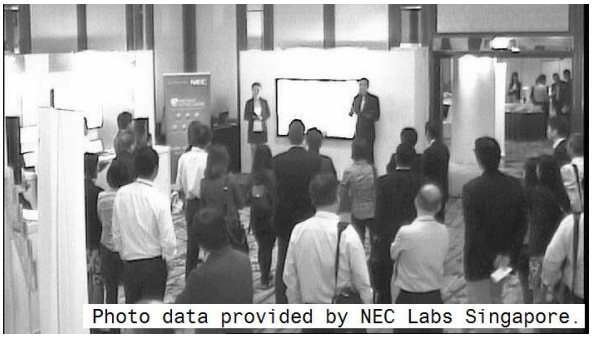

Figure 2 Example of crowded video

3.2 Flexibility

In wide-area hub, video will be analyzed to find anomaly in the field (A_k in the Figure 1). Table 1 shows some example of analysis use case and camera configurations. Furthermore, the camera configuration, such as location, angle and zooming is determined according to a video analysis in wide-area hub (A_k). Therefore, different compression functions f_k in local hubs will be required for different video analysis (A_k) and camera configurations.

For example, in case of use case (a) "find person" with low depression angle and zoomed camera, the face detection may be suitable as labeling algorism because user would like to see video in which persons are captured and it is good enough to identify who they are. On the other hand, in case of use case (b) "find suspicious behavior" with middle depression angle from far camera, motion detection can be suitable as labeling algorism to filter out the video segments in which no one is captured because the suspicious behavior, such as climbing a fence, tend to occur in a place where no one around and the video tend to be sparse. For use case (c) "find unusual crowd behavior" with middle depression angle from far camera, some new technology will be needed as we discussed in the forgoing section.

Table 1 Example of use case and camera configurations

Use case	Camera configuration
(a) Find person (ex: Person of Interest detection)	The camera is configured with low depression angle and zoomed to capture personal features (face, clothing, etc.). The video tend to be crowded.
(b) Find suspicious behavior (ex: intruder detection)	The camera is configured with middle depression angle from far to capture the movement of person. The video tend to be sparse.
(c) Find unusual crowd behavior (ex: crowd congestion detection)	The camera is configured with middle depression angle from far to capture the movement of people. The video tend to be crowded.

Therefore, the system should be capable of flexibility which allows us to apply suitable labeling algorisms for variety of video analysis in wide-area hub and camera configurations combinations.

4 SYSTEM ARCHITECTUERE

4.1 Approach

We took following approach to tackle the two challenges discussed above: we applied crowd congestion detection algorism as labeling algorism for crowded videos, and we provide a frame work for content base quality control system with labeling engine plug and play.

The crowd congestion detection is selected as an example of supporting crowded video. We tested the effect of this technology on data compression f_k.

We designed a framework using a middleware for large-scale video surveillance system called Analysis Control Middleware (ASCOT) [8]. We modularized the labeling algorisms as labeling engines and ASCOT allows us to change or add new labeling engines so that system integrator can select best combination.

4.2 Architecture

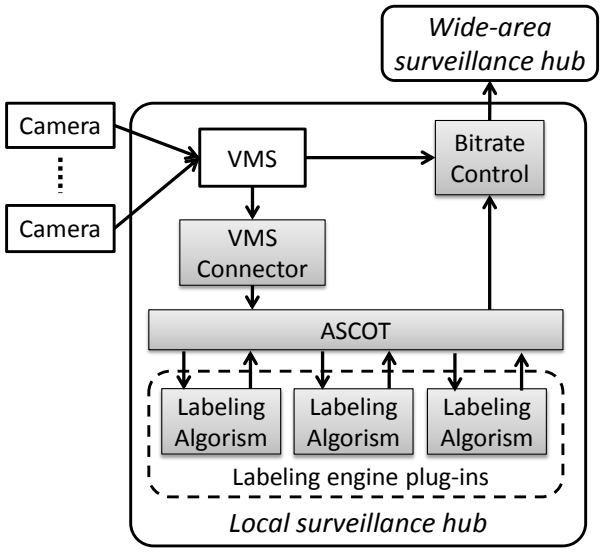

Figure 3 System architecture

The system includes connector to local video managing software (VMS), labeling algorism engine plug-ins, bitrate control and ASCOT which connects these modules (Figure 3).

The system is designed to be able to insert between existing VMS in local hub and wide-area hub. The VMS connector will receive video streams from cameras for the labeling engines. System integrator can select or add suitable labeling engines which are provided as plug-ins for each video streams on ASCOT. These labeling engines output score of importance and the bitrate control module will change the resolution and frame rate of the streams based on the scores. User can change or add new algorism flexibly because VMS connector, labeling engines and bitrate control are modularized with ASCOT API.

5 EVALUATION

5.1 Experimental Methodology

We developed a system to evaluate followings with a case study application.

(1) Estimated data reduction by face detection for corridor video.
(2) Estimated data reduction by congestion detection for entrance video.
(3) Estimated data reduction by motion detection for both of videos as a base line.
(4) Flexibility by changing these three labeling algorisms for each camera.

As a case study application, we selected important premises surveillance which has cameras in corridors for person of interest detection and cameras in entrances for unusual crowd behavior detection. We used real surveillance videos captured in a premise in Singapore as shown in Table 2.

To estimate data reduction, we used following bitrate change rules as the bitrate control:

(1) Send the original stream for the video segments which are marked as important.
(2) Reduce the bitrate to 500kbps by reducing fps and resolution of the video segments which are marked as not important.

Table 2 Video feeds used for the evaluation

Video	Description	Quality
Corridor video	The camera is configured with low depression angle and zoomed to capture personal features. (20 hours: 4am-0pm).	960x540, 5fps, mpeg2, 2.4Mbps
Entrance video	The camera is configured with middle depression angle from far to capture the movement of people. (20 hours: 4am-0pm)	640x480, 5fps, mpeg2, 2.0Mbps

5.2 Evaluation Results

5.2.1 Estimated Data Reduction

The data reduction was estimated for both corridor video and entrance video. The ratio of the frames marked as unimportant by labeling algorisms and estimated data reduction are shown in the Figure 4

For corridor video, the face detection showed reduction of 59% frames results in 46% data reduction. Motion detection showed reduction of 36% frames results in 27% data reduction for the same video footage.

For entrance video, the congestion detection showed reduction of 71% frames results in 41% data reduction with 0.4 congestion level which is same density as Figure 2 where around 20 people in the view. Motion detection showed reduction of 23% frames results in 6% data reduction for the same video footage.

For both cases, the accuracy of video analysis (A_k), such as face matching and overcrowd detection, will remains same because the original streams are sent for the video segments marked important.

The data reduction by the motion detection is concentrated in the morning and night time in which nobody in the premises for both corridor and entrance cases. This leads to the 25–74% lower data reduction compare to frame reduction because the video codec efficiently compress the video for these period as discussed in the next section.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

On the other hand, face detection was able to filter out the scenes in which only parts of persons or their belongings was captured. Congestion detection was able to filter out the scenes in which people can walk without bumping each other that can be considered as normal congestion. This leads to that face detection and crowd detection can keep the reduction rate high compare to motion detection.

Therefore, face detection and crowd detection can reduce 1.7 times and 6.8 more data respectively compare to motion detection as shown in Figure 4.

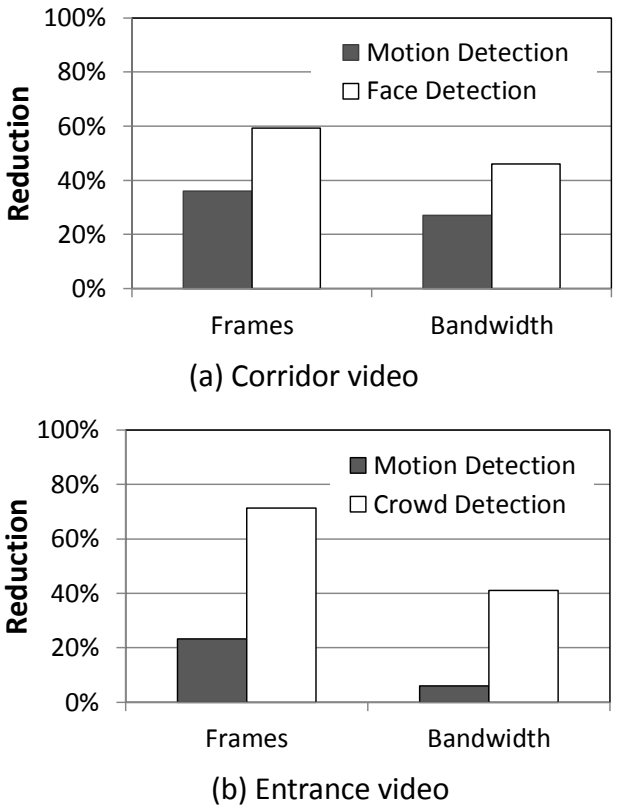

Figure 4 The estimated data reduction

The data reduction for each congestion level is also shown in Figure 5. The threshold of the congestion level to mark a frame as important or unimportant will be decided in a subjective manner based on the use case of the video feed. As shown in the graph, the congestion detection can reduce 6-73% bandwidth, which is larger than the data reduction by motion detection (6%). Therefore we can conclude that the congestion detection can be a potent candidate for the data reduction for crowded videos.

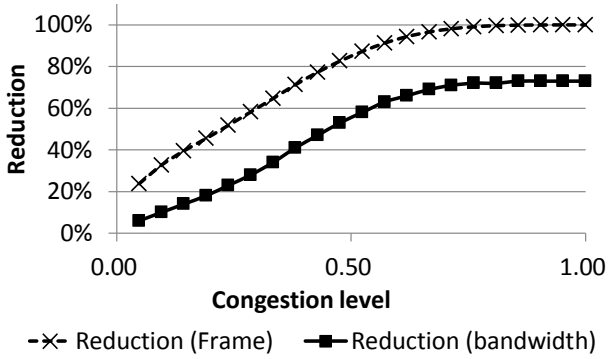

Figure 5 The data reduction vs congestion level for entrance video

5.2.2 Correlations between Labeling Results and the MPEG2 Bitrates

We analyzed the correlations between labeling results and MPEG2 bitrates of the video segments to clarify the reason why the data reduction by motion detection is less than other labeling algorisms. The results are shown in the Figure 6 and Figure 7. In these graphs, the numbers of labeled frames are counted for each bitrate range. For the crowd detection, we used a congestion level 0.4 as the threshold.

The results show that face detection and crowd detection can remove the frames which having higher bitrate compare to motion detection. For the entrance video, the most of the frames marked “with motion” have bitrate range of 1.05 to 3.15 Mbps (Figure 7). On the other hand, the frames marked “crowded” have higher bitrate range 2.1 to 3.5Mbps (Figure 7). The same is true for corridor video (Figure 6). The frames removed by motion detection tend to have low bitrate because the codecs such like mpeg2 and H.264 are designed to compress the data by using differences between frames [9].

In addition, we can see same trend even if the threshold of crowd congestion level is changed because there is correlation between the bitrate and congestion level as shown in Figure 8.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

5.2.3 Flexibility of the System

The system successfully applied three different labeling algorisms, i.e. motion detection, face detection and congestion detection. They showed effective data reduction for different use cases and camera configurations. In addition, the customizability of the labeling algorisms on ASCOT is self-evident because of engine plug-and-play capability by standardized API called Exec APIs as discussed in [8].

Therefore, we conclude that the system has flexibility to support a broad range of the use cases and camera configurations.

5.3 Considerations

The results of the evaluation showed that the content base quality control can reduce the data more than 40% by applying suitable labeling algorism for each use case. However, we need to consider countermeasures for the situation in which band width reaches limit because many video segments are marked important in the same time.

One of the candidates is combination of the content base quality control and traditional bandwidth management which based on the network status monitoring. In this case, the content base quality control takes care of the

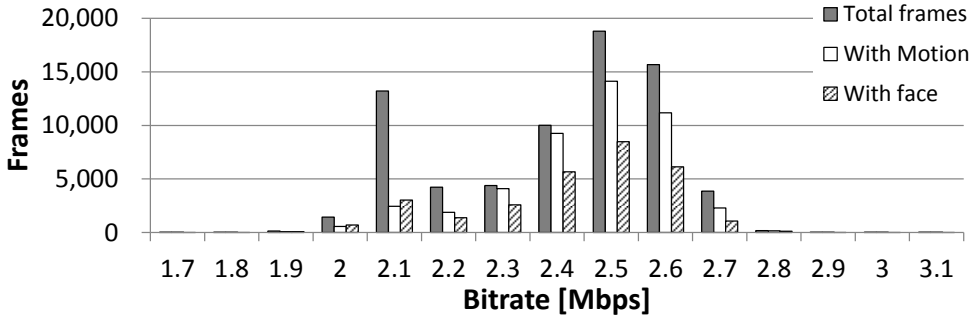

Figure 6 Bitrate vs analytic results for corridor video

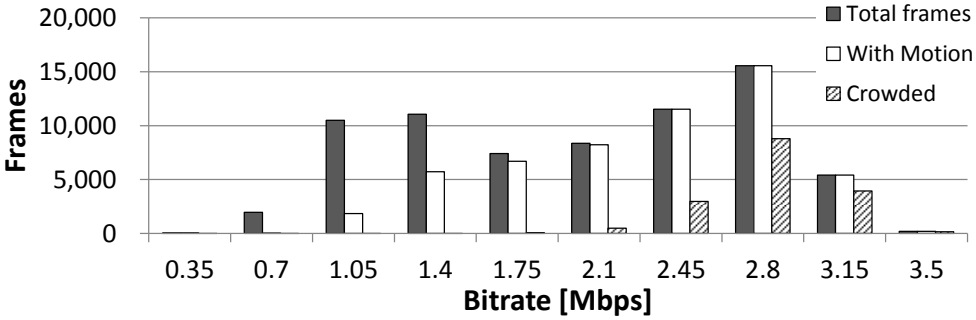

Figure 7 Bitrate vs analytic results for entrance video

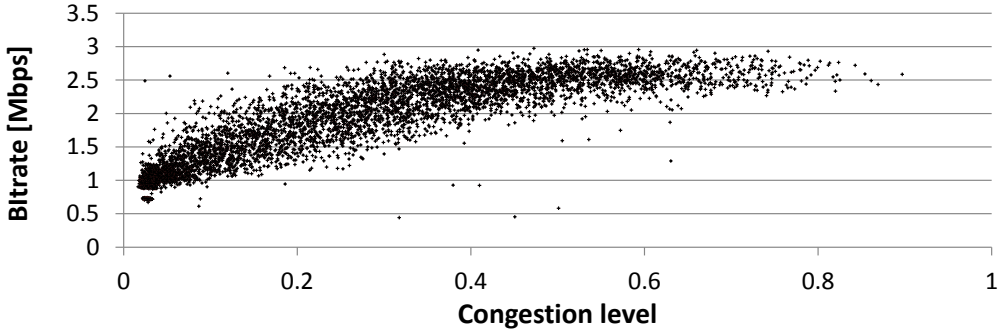

Figure 8 Bitrate distribution vs congestion detection for entrance video

data compression during normal operation and if the bandwidth reached the limitation, traditional bandwidth management interpose to prevent the network congestion. To ensure the surveillance quality, some prioritization for the traditional bandwidth management will be also needed based on the use case importance for the surveillance operators. This kind of enhancement can be implemented as the bitrate control module, which is modularized using ASCOT API, in the proposed video quality control system.

To design and evaluate this kind of solution, we need to clarify following two things.

(1) The patterns of the labeling results: How often the video segments from different cameras are marked important in the same time? Is there any pattern on that?
(2) Survey on surveillance use cases: Is there any priority between use cases? Does the priority change by the context such as occurrence of incidents?

As a future work, we will apply the system to multiple real surveillance systems to answer these questions to design and evaluate the solution.

6 CONCLUSION AND FUTURE WORK

This paper focused on the data reduction to realize wide-area surveillance systems. We highlighted two major challenges we face in real surveillance system, i.e. support of crowded videos, and flexibility for supporting broad range of use cases and camera configurations. We proposed use of congestion detection for data reduction and configurable framework using a middleware called ASCOT for the system.

We evaluated them with two real surveillance video footages, one from corridor and another one from entrance taken in important premises. Crowd congestion detection reduced 41% of data for the entrance video footage, and face detection reduced 46% of data for the corridor video footage where motion detection reduced only 6% and 27% respectively. Furthermore, we confirmed that the system can support new labeling algorisms and flexibly combine them to meet the requirement of each video streams.

As a future work, we will investigate and evaluate the countermeasures for the situation in which many video segments are marked important by applying the system to real surveillance systems. Through these activities, we hope to make it possible to develop intelligent city surveillance systems to make cities safer.

7 REFERENCES

[1] March Networks, “Singapore MRT Moves Ahead on CC TV System Expansion”, March Networks News October 2008 – Transportation edition, 2008, pp. 14-15

[2] Cozzolino, A., Flammini, F., Galli, V., Lamberti, M., Poggi, G. and Pragliola, C., Evaluating the Effects of MJPEG Compression on Motion Tracking in Metro Railway Surveillance, Proceedings of the 14th International Conference on Advanced Concepts for Intelligent Vision Systems, Springer-Verlag, 2012, pp. 142-154

[3] Korshunov, P. and Ooi, W. T., “Critical Video Quality for Distributed Automated Video Surveillance”, Proceedings of the 13th Annual ACM International Conference on Multimedia ACM, 2005, pp. 151-160

[4]

[5] Pillai, P., Ke, Y. and Campbell, J., “Multi-fidelity Storage”, Proceedings of the ACM 2Nd International Workshop on Video Surveillance & Sensor Networks ACM, 2004, pp. 72-79

[6] Hengstler, S., Prashanth, D., Fong, S. and Aghajan, H., ”MeshEye: A Hybrid-resolution Smart Camera Mote for Applications in Distributed Intelligent Surveillance”, Proceedings of the 6th International Conference on Information Processing in Sensor Networks, 2007, pp. 360-369

[7] Chen, K., Loy, C. C., Gong, S. and Xiang, T., “Feature Mining for Localised Crowd Counting”, Proceedings of the British Machine Vision Conference BMVA Press, 2012, pp. 21.1-21.11

[8] Rahmalan, H., Nixon, M. and Carter, J., “On Crowd Density Estimation for Surveillance”, Crime and Security, 2006. The Institution of Engineering and Technology Conference on, 2006, pp. 540-545

[9] Arikuma, T., Koyama, K., Kitano, T., Shiraishi, N., Nagai, Y. and Kawamata, T., “Analysis control middleware for large-scale video surveillance”, 2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance, IEEE Computer Society, 2013, pp. 294-299

[10] Wiegand, T., Sullivan, G., Bjontegaard, G. and Luthra, A., “Overview of the H.264/AVC video coding standard”, Circuits and Systems for Video Technology, IEEE Transactions on, 2003, Vol. 13(7), pp. 560-576