Rate–Accuracy Tradeoff in Automated, Distributed Video Surveillance Systems

Pavel Korshunov
School of Computing
National University of Singapore
pavelkor@comp.nus.edu.sg

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Video Analysis; C.2.4 [**Distributed Systems**]: Distributed Applications

General Terms

Experimentation, Measurement, Performance

Keywords

Video Surveillance, Video Analysis, Video Features, Rate–Accuracy Tradeoff

1. INTRODUCTION

For the past years video analysis algorithms are attracting more research interest and becoming commonly used in various multimedia applications. The increasing accuracy of these algorithms encourages their use for navigating robots, indexing and searching in video and image databases. One of the most popular areas for their application is video surveillance where video analysis algorithms are used for detection, tracking and analysis of people, objects and events. Since video analysis algorithms operate over large amounts of data, such as videos and images, they impose high system requirements on storage and computational resources, network bandwidth, and power for battery–powered devices. Despite this fact, surprisingly little attention was paid to the resources minimization problem accompanying the use of video analysis algorithms.

In this paper we focus on the problem of system resources reduction in automated large–scale distributed video surveillance systems, which use video analysis algorithms for automatic monitoring. According to the study by Wu *et al.* [6], suspicious events are rare in video surveillance. Therefore, most of the time surveillance video is monitored by various video analysis algorithms instead of human observers. However, existing video compression algorithms and video presentation standards are tuned to the human visual system. It allows us to take the advantage of different, compared to the human vision, quality requirements of video analysis algorithms. Therefore, in this work, we study the effect of compression, resizing, dropping of frames in the video, and other video adaptations on the accuracy of video analysis algorithms.

Through extensive experiments with face detection and face tracking algorithms we found that these algorithms can sustain a significant degradation in the input video quality without decrease in their accuracy. Therefore, the tradeoff between the video quality and the performance accuracy of algorithms exists. The face detection algorithm even shows an increase in the detection accuracy if digital zooming is applied without changing the video bit rate.

The above results encourage us to consider a general system resource optimization problem in context of different video adaptations and various video analysis algorithms instead of the bit rate minimization problem for the specific algorithm. This problem is based on the *rate–accuracy tradeoff*, since video adaptations effectively reduce video bit rate and algorithms are evaluated by their accuracy. The benefit of studying the rate–accuracy tradeoff is twofold. First, it helps in solving the rate minimization/accuracy maximization problem. Second, it can give insights on developing new video compression algorithms tuned for greater performance of video analysis algorithms and provide guidelines on building video analysis algorithms aimed to have higher accuracy for the video with reduced quality.

The obvious way to find the rate–accuracy tradeoff for any video analysis algorithm is to run many offline experiments with various adaptations on input videos. Such an approach, however, is not practically feasible because of the large number of video analysis algorithms. We propose, therefore, to study the effect of video adaptations on video features used by algorithms instead of simply measuring the accuracy of every algorithm with respect to these adaptations. We base it on the hypothesis that most of video analysis algorithms rely their video analysis on a closed set of video features. For example, the face tracking algorithm is based on color histograms of tracked objects. We believe that the number of such video features is relatively small and is almost independent from the diverse collection of video analysis algorithms. Video features can be also determined from the video directly. Studying relationships between video features and video adaptations helps us to estimate the rate–accuracy tradeoff for an arbitrary video analysis algorithm. It might even provide guidelines on how to design a new video analysis algorithm with good rate–accuracy characteristics.

We propose a formal framework describing the rate–accuracy tradeoff, which is analogous to utility–based adaptation [2] and rate–distortion optimization [3] frameworks. The rate–accuracy framework suggests a method for deriving the tradeoff for a particular video analysis algorithm by using the cumulative effect of video adaptations on corresponding video features. We developed a distributed prototype video surveillance system showing practical benefits of proposed rate–accuracy framework.

In Section 2 we present our experimental results when studying the rate–accuracy tradeoff of face detection and face tracking algorithms. In Section 3 we shift the focus to studying the effect of video adaptations on video features used by video analysis algorithms, and present a preliminary classification of video features for detection and tracking algorithms. Our vision on the future de-

MM'06, October 23–27, 2006, Santa Barbara, California, USA.
ACM 1-59593-447-2/06/0010.

velopment of the formal rate–accuracy framework and how we can apply it in our prototype video surveillance system is presented in Section 4. Lastly, we give a brief overview on the related work in Section 5.

2. RATE–ACCURACY TRADEOFF

In this section we experimentally verify the hypothesis that the bit rate of video, which is monitored exclusively by video analysis algorithms, can be reduced without decreasing accuracy of algorithms. We conducted a set of experiments that study the effect of reduced video quality, i.e. video bit rate, on the accuracy of two typical video analysis algorithms. We used CAMSHIFT face tracking algorithm and Viola-Jones face detection algorithms implemented in OpenCV library. As a testing dataset we used MIT/CMU collection of images for face detection, some parts of movies for face tracking and different lab videos for both. In our experiments we considered three video adaptations that affect video bit rate: i) Reduction of compression quantizer, i.e. changing SNR video quality; ii) Frames dropping, i.e. changing temporal video quality; iii) Bicubic re-scaling, i.e. changing spatial video quality. For face detection we were changing SNR quality and spatial quality; and for face tracking, SNR quality and temporal quality. The temporal quality affects the face detection algorithm mainly as a filter that reduces false positive. More detailed description of experiments, their implications and practical use of the results can be found in [5].

Our experiments show that accuracy of the algorithm can sustain large drops in video qualities which result in significant reduction of the video bit rate. The typical rate–accuracy curve has a behavior that is similar to the one presented in Figure 1. The curve shows a *sweet spot* that represents the quality to which video can be reduced without affecting the resulting accuracy of an algorithm. This result demonstrates that even for quite complicated video analysis algorithms such as face detection and face tracking the video bit rate can be reduced drastically. At the same time the false positive does not show a noticeable growth when quality is decreased. In our experiments with the prototype implementation of video surveillance system we obtained 29 times reduction in bit rate for face detection and 16 times reduction for face tracking.

Experiments with Bicubic zooming adaptation show that even an improvement can be achieved in the accuracy of face detection algorithm if input images are pre-scaled up using Bicubic algorithm. It shows that at the expense of slight increase in false positive, Bicubic zooming gives the face detection algorithm ability to detect more of small faces. Experiments in the lab environment show that when zooming was applied the face detection algorithm could detect faces that previously fell under limits of its detectable face size.

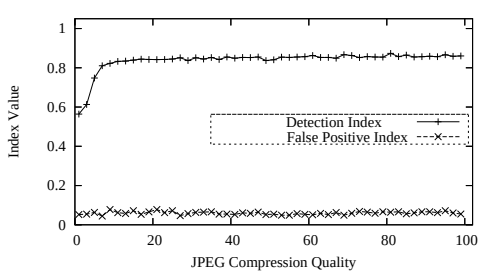

Figure 1: Accuracy of Face Detection Algorithm vs. JPEG Compression Quality.

All experiments presented above also show that various video adaptations have distinctive impacts on the rate–accuracy tradeoff. While Bicubic zooming can even improve the accuracy of face detection despite the unchanged video bit rate, the frame dropping adaptation affects the accuracy of the algorithm only indirectly, helping to reduce the false positive [5]. Other common video adaptations that can be considered include: low or high pass filtering, changes in luminance, drop in DCT coefficients, optical zooming, etc.

3. VIDEO FEATURES

The straightforward way to find out the rate–accuracy tradeoff for a new video analysis algorithm is to run the set of experiments similar to the one presented in Section 2. However, in general there can be many different video adaptations affecting the video quality. For instance, SNR quality can be reduced by quantizing the video, dropping DCT coefficients or using different kind of compression algorithm like JPEG2000, etc. In the same time the number of existing and constantly updated video analysis algorithms is very large. Moreover, video analysis algorithms are not always available for before ahead offline experiments. One way to obtain or at least estimate the rate–accuracy tradeoff for an arbitrary video analysis algorithm is to categorize algorithms into groups, thus, reducing the set size needed to be considered. The space of video adaptations can also be reduced to several main categories. We noted that both algorithms and adaptations operate on video. In fact, the video is a binding point which defines the quintessence of the relationship between a video analysis algorithm and a video adaptation (see Figure 2).

Therefore, we suggest studying the video content which affects both sides of the algorithm–adaptation relationship. We propose the hypothesis that most of video analysis algorithms depend on a closed set of video characteristics and properties that we call *video features*. Such features are important for the performance of video analysis algorithms and are affected by video adaptations in different ways. For example, the face tracking algorithm relies on the color histogram for the objects of limited spatial size. We aim to identify the common set of video features that are used by typical video analysis algorithms.

In our preliminary studies we try to identify types of video features used by arbitrary detection and tracking algorithms. We suggest that video features can have different video properties, i.e. temporal, spatial, histogram, color, etc. Examples of video features can include: sizes of objects, edges, color histograms, the speed of a moving object, haar features, etc. In this paper we suggest the preliminary classification of temporal and spatial video features used by tracking and detection algorithms.

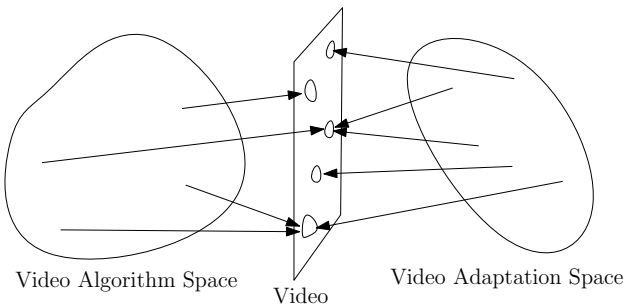

Figure 2: The Relationship between Video Analysis Algorithms and Video Adaptations.

Temporal features can be categorized as following: (i) High ratio of the speed to the size of tracked object. It appears in eye tracking, face tracking using eyes, nose, and corners of lips, airplanes tracking, and some cases of car tracking. (ii) Low ratio of object's speed to its size. This type is present in silhouette tracking, face tracking based on color histogram, and car tracking in car park scenario. Spatial features can be in one of these types: (i) Small size. Such features are used in eye tracking, face recognition, finger prints analysis, and pins and whole position detection in machinery. (ii) Rough features. Used in face detection based on haar features, briefcase detection, and the identification of a vehicle type. (iii) Borders and edges. Used in silhouette detection, barcode reading, the vehicle type identification, the building structure analysis, and bones detection in X-ray. (iv) Blob features. Used in car park car tracking, face tracking based on color histogram, and general foreground object detection.

In our future work we aim to identify the closed set of video features that are used by video analysis algorithms in video surveillance. This set should include such video features that are most influential in terms of the rate–accuracy tradeoff. For instance, most of object detection algorithms have a limit on the minimal detectable object size. This feature has a strong implication on the rate–accuracy tradeoff of a detection algorithm and is affected by the zooming adaptation.

4. RATE–ACCURACY FRAMEWORK

The main objective of the rate–accuracy framework is to give general guidelines on how to estimate the rate–accuracy tradeoff for an arbitrary video analysis algorithm. To answer this question we first replace the video analysis algorithm with its video features as described in Section 3. Video features can be characterized by different video properties, i.e. temporal, spatial, color, etc. Each characteristic is affected by only a few video adaptations, for example, the face size is affected by spatial resizing of the frame but not by DCT coefficient dropping. This observation demonstrates that the set of relationships between video adaptations and video features is practically not very large. Among the remaining feature–adaptation dependencies, some can be described analytically, for example the effect of zooming on the face size. Other dependencies can be analyzed experimentally by varying the degree of adaptation applied to the video with studied feature and measuring changes that the feature undergo. However, it is not always clear how to measure the adaptation effect on features, for example, how to quantitatively describe changes in color histogram caused by JPEG compression of the image. The problem of clearly defining every relationship that binds video feature and video adaptation spaces is the subject of our future work.

Lets assume that we can find a function or heuristic describing the effect of a video adaptation on every video feature used by the video analysis algorithm. The problem then is to merge these functions into some formula that can be used as an estimation of the desired rate–accuracy tradeoff. It can be done by combining the normalized forms of these functions using, for instance, some linear weighted function with variable weights corresponding to different feature–adaptation relationships. The assignment of weights is not clear as well, but looking at the face detection example we can notice that SNR affects the detection accuracy only below the sweet spot, while spatial resizing has almost linear and continuous impact on the accuracy. The resulted weighted function can serve as a heuristic estimation on the rate–accuracy tradeoff. The practical benefit of this approach should be verified using several typical video analysis algorithms by comparing obtained estimations with actual values of rate–accuracy functions.

To analyze the rate–accuracy framework in practice we built the prototype video surveillance system, which currently uses face detection and face tracking algorithms. We proposed to use a distributed architecture for video surveillance system consisting of video sources, processing proxies and monitoring stations. Processing proxies in such system run video analysis algorithms which filter the video coming from video sources. Therefore, the bit rate of the video streamed at a link between a proxy and a video camera can be reduced using the rate–accuracy tradeoff. Only in rare occurrences of suspicious events it is required to stream the full quality video from the video source via proxy to the monitor for the inspection by the human observer.

5. RELATED WORK

The concept and framework of the rate–accuracy function is similar to the conceptual framework based on utility function described by Chang *et al.* [2]. The notion of rate–distortion is generalized to human–oriented utility function which describes the relationship between different video–adaptations and different video resources defined as video entities, i.e. frames, pixels, etc. Such generalization allows solving utility optimization problems with constrained video bit rate. Authors do not consider how to define the utility function when many different adaptations are applied to the video. In our work, the utility function, i.e. accuracy also can be obtained experimentally from a video analysis algorithm. The problem is that instead of a single human observer we have many video analysis algorithms showing different performance accuracy for different video adaptations.

Comparing our work with current research in video surveillance systems we noticed that most works aim at developing robust computer video analysis and classification algorithms [4]. Other works on video surveillance systems do not pay attention to system issues such as efficiency and scalability proposing centralized system architectures and assuming the availability of network resources [1].

6. REFERENCES

[1] E. Y. Chang, Y.-F. Wang, and I.-J. Wang. Toward building a robust and intelligent video surveillance system: a case study. In *Proceedings of the IEEE International Conference on Multimedia and Expo, ICME'04*, pages 1391–1394, Taipei, Taiwan, June 2004.

[2] S.-F. Chang and A. Vetro. Video adaptation: Concepts, technologies and open issues. *Proceedings of the IEEE*, 93:148–158, Jan. 2005.

[3] A. Eleftheriadis and D. Anastassiou. Constrained and general dynamic rate shaping of compressed digital video. In *Proceedings of the IEEE International Conference on Image Processing, ICIP'95*, pages 396–399, Washington, DC, USA, Oct. 1995.

[4] V. Kettnaker and R. Zabih. Bayesian multi-camera surveillance. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'99*, volume 2, pages 253–259, Fort Collins, Colorado, USA, June 1999.

[5] P. Korshunov and W. T. Ooi. Critical video quality for distributed automated video surveillance. In *Proceedings of the 13th ACM International Conference on Multimedia, ACMMM'05*, pages 151–160, Singapore, Nov. 2005.

[6] Y. Wu, L. Jiao, G. Wu, E. Chang, and Y.-F. Wang. Invariant feature extraction and biased statistical inference for video surveillance. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS'03*, pages 284–289, Miami, FL, July 2003.