# SURVEILLANCE VIDEO FOR MOBILE DEVICES

*Olivier Steiger, Touradj Ebrahimi*

Signal Processing Institute
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
{olivier.steiger,touradj.ebrahimi}@epfl.ch

*Andrea Cavallaro*

Multimedia and Vision Lab
Queen Mary, University of London
Mile End Road, London E1 4NS, UK
andrea.cavallaro@elec.qmul.ac.uk

## ABSTRACT

In this paper, we present a video encoding scheme that uses object-based adaptation to deliver surveillance video to mobile devices. The method relies on a set of complementary video adaptation strategies and generates content that matches various appliance and network resources. Prior to encoding, some of the adaptation strategies exploit video object segmentation and selective filtering in order to improve the perceived quality. Moreover, object segmentation enables the generation of automatic summaries and of simplified versions of the monitored scene. The performance of individual adaptation strategies is assessed using an objective video quality metric, which is also used to select the strategy that provides maximum value for the user under a given set of constraints. We demonstrate the effectiveness of the scheme on standard surveillance test sequences and realistic mobile client resource profiles.

## 1. INTRODUCTION

The problem of remote visual surveillance of unattended environments has received growing attention in recent years. But whereas event monitoring is still mostly performed by human operators located in a fixed surveillance room, there is an increasing demand for the delivery of surveillance video to mobile devices as well. The latter notably enables surveillance personnel to monitor a critical situation without interruption even while shifting to the intervention place. Moreover, it permits the surveillance of private ground and vacant homes using cellular phones and PDAs. However, reliable remote surveillance requires the quality of the delivered video to be optimal despite the limitations resulting from small display sizes and restricted processing capabilities of mobile devices. In addition to this, one must cope with the limited bandwidth and time-varying conditions of wireless transmission channels.

Traditionally, scalable video coding [1] and content-blind transcoding [2] have been used to adapt video to the restricted capabilities of mobile terminals and networks. However, scalable video coding requires specific decoding capabilities to access individual quality or resolution layers. Moreover, scalable video streams are not optimal in terms of the required bandwidth. The above problems are solved by using transcoding, where the video is adapted to the capabilities of the receiver at the encoder's side. However, traditional transcoding techniques (content-blind techniques) are generally not optimal in terms of perceptual quality. Thus, recent transcoding methods (content-based techniques) make use of content characteristics in order to minimize the degradation of important image regions. In particular, object-based transcoding considers the usage of video objects as transcoding entities. That is, foreground objects are encoded at a higher quality level or resolution than less important regions [3, 4]. While the works in [3, 4] resort to object-based encoders (e.g., MPEG–4) to code different image regions individually, we present in this paper a method that exploits an object-based representation in a traditional frame-based encoding framework (e.g., MPEG–1). The rationale behind this choice is to enable the use of advanced functionalities with standard decoders available for consumer devices. Also, the additional knowledge provided by object-based analysis can further be exploited to meet the restricted capabilities of mobile devices.

The remainder of this paper is organized as follows. In Section 2, we discuss the generation and delivery of video that matches the resources of mobile devices in an optimal way. In Section 3, results obtained with real surveillance sequences are presented and discussed. Finally, the conclusions of our work are drawn in Section 4.

## 2. DELIVERY OF SURVEILLANCE VIDEO

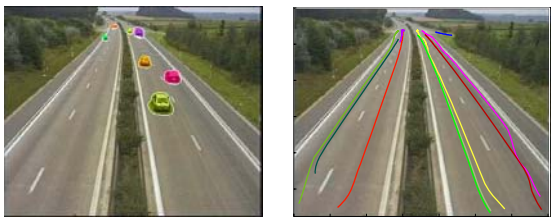
Adaptive delivery ensures that the delivered video matches the limited capabilities of mobile appliances in an optimal way. This is achieved by transforming the video using a number of complementary adaptation strategies, and by selecting the strategy that provides most perceptual quality for the end user.

(a) (b) (c) (d) (e)

**Fig. 1**. Background simplification for compression improvement and objects enhancement. (a) The original background from the *Highway* sequence is replaced by a static background shot. (b) The original background is lowpass-filtered. (c) Each background macroblock is replaced by its DC value. (d) An edge image is used instead of the original background. (e) Background areas are set to a constant value.

### 2.1. Video adaptation strategies

The uncompressed input video is first transformed using one out of the following adaptation strategies: *coded original*; *spatial resolution reduction*; *semantic prefiltering*. The *coded original* is simply obtained by encoding the input video using a frame-based encoder, such as MPEG–1. *Spatial resolution reduction* can further be applied prior to the coding in order to reduce the transmission bandwidth. *Semantic prefiltering* aims at mimicking the way humans treat visual information in order to improve the compression ratio of image and video coders, and to enhance relevant objects [5]. To enable semantic prefiltering, image areas that observers are looking at (foreground) need to be separated from areas that are not expected to attract the attention of a viewer (background) by means of video object segmentation [6]. The overall image quality is then improved by simplifying the background in order to improve the quality (i.e. the associated bit allocation) of the foreground. This is achieved by replacing the original background by a static background shot, by lowpass-filtering the background, or by replacing each background macroblocks by its DC value (Fig. 1(a)-(c)). Alternatively, superfluous visual details may be removed from the background to enhance relevant objects (Fig. 1(d)-(e)).



(a) (b)

**Fig. 2**. Use of object segmentation to meet restricted device capabilities. (a) Relevant objects are put in a conspicuous situation on small displays. (b) Surveillance video is summarized in a single image.

The additional knowledge provided by object segmentation can further be exploited to meet the restricted capabilities of mobile devices. In Fig. 2(a), relevant video objects have been put in a conspicuous situation by means of colored blobs. This is particularly useful on small displays. In Fig. 2(b), the surveillance video has been summarized in a single frame by plotting the trajectories of semantic video objects on top of a static background shot. This can be used to convey the meaning of the filmed scene when video capabilities are not available.

The perceptual quality resulting from the individual adaptation strategies is then evaluated by means of objective evaluation. An objective video distortion measure that emulates human judgement needs to account for different image areas and for their relevance to the observer. This aspect can be considered with the traditional Mean Squared Error (MSE) by weighting different image areas according to their semantics. This leads to the *semantic mean squared error*, SMSE, defined as [5]:

$$\mathrm{SMSE} = \sum_{k=1}^{N} \frac{w_k}{|C_k|} \sum_{(i,j)\in C_k} d^2(i,j), \qquad (1)$$

where $N$ is the number of classes and $w_k$ the weight of class $k$. Class weights are chosen depending on the semantics, with $w_k \geq 0, \forall k = 1, \ldots, N$ and $\sum_{i=1}^{N} w_k = 1$. $C_k$ is the set of pixels belonging to the object class $k$, and $|C_k|$ is its cardinality. The error $d(i,j)$ between the original image $I_O$ and the distorted image $I_D$ in Eq. (1) is the pixel-wise color distance. The color distance is computed in the 1976 CIE *Lab* color space in order to consider perceptually uniform color distances with the Euclidean norm. The final quality evaluation metric, the *semantic peak signal-to-noise ratio* SPSNR, uses SMSE instead of MSE as compared to PSNR. When the classes are foreground and background, then $N = 2$ in Eq. (1), and $w_f$ is the foreground weight. The background weight is thus $(1 - w_f)$.

### 2.2. Strategy selection

Strategy selection is at last needed to work out the adaptation strategy that provides most perceptual quality for the end user, considering the individual resources of the connected client (i.e., appliance, network).

Specifically, let $A_i$ be some original item, e.g., a video. The adapted version $M_{ijk}$ is computed by transcoding $A_i$ using the adaptation operator $O_j$ at resources $k$. Each adaptation operator $O_j$ implements an adaptation strategy in Section 2.1. The perceptual quality of $M_{ijk}$ resulting from the adaptation is denoted by $Q(M_{ijk})$. Let us furthermore define the item resource vector for the item $M_{ijk}$ as $\mathbf{R}(M_{ijk}) = \left(R(M_{ijk})^1, R(M_{ijk})^2, \ldots, R(M_{ijk})^r\right)^T$, where $r$ is the number of different resources that have to be considered (e.g., bitrate, resolution, coding format, etc.). Similarly, the client resource vector is denoted by $\mathbf{R}_{\text{client}} = \left(R^1_{\text{client}}, R^2_{\text{client}}, \ldots, R^r_{\text{client}}\right)^T$.

The selection of the optimal adaptation strategy can then be formalized by the following resource allocation problem:

**Problem 1** *For item $A_i$, find the adapted version $M_{ijk}$ that has maximum quality $Q(M_{ijk})$ such that item resources $\mathbf{R}(M_{ijk})$ do not exceed client resources $\mathbf{R}_{\text{client}}$:*

$$\max_{j,k}\left\{Q(M_{ijk})\right\} \quad \textit{such that} \qquad (2)$$
$$R^n(M_{ijk}) \leqslant R^n_{\text{client}} \quad \textit{for all } 1 \leqslant n \leqslant r$$

In order to solve Problem 1, we define a number of *anchor nodes* $\left(O_j, \mathbf{R}(M_{ijk}), V(M_{ijk})\right)$. An anchor node expresses the quality $Q(M_{ijk})$ resulting from applying adaptation operator $O_j$ at resources $\mathbf{R}(M_{ijk})$. We further fit a polynomial *quality function* (QF) to the anchor nodes of each adaptation operator. The quality function matrix for $A_i$ is denoted as

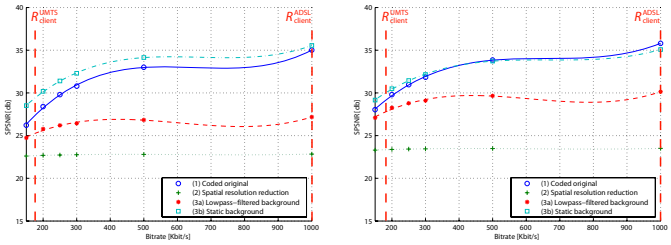
$$\mathbf{F}^{\text{QF}}_i = \left(\mathbf{f}^{\text{QF}}_{i1}, \mathbf{f}^{\text{QF}}_{i2}, \ldots, \mathbf{f}^{\text{QF}}_{iJ}\right)^T = \begin{pmatrix} a_{i1,1} & a_{i1,2} & \ldots & a_{i1,p} \\ a_{i2,1} & a_{i2,2} & \ldots & a_{i2,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{iJ,1} & a_{iJ,2} & \ldots & a_{iJ,p} \end{pmatrix}, \qquad (3)$$

where $a_{ij,k}$ are the coefficients of the order $p-1$ polynomial quality function, $\mathbf{f}^{\text{QF}}_{ij}$. $J$ is the number of adaptation operators.

The solution to Problem 1 is then given by the QF that has maximum quality $\max_{j,\mathbf{R}}\left\{\mathbf{f}^{QF}_{ij}(\mathbf{R})\right\}$ such that $\mathbf{R} \leqslant \mathbf{R}_{\text{client}}$. In the particular case where all QFs are monotonically increasing, the solution is located at $\mathbf{R} = \mathbf{R}_{\text{client}}$.

## 3. RESULTS

In this section, the proposed adaptive delivery framework is tested with surveillance sequences and realistic client resource profiles. In particular, the mechanism discussed in Section 2.2 is used to select the adaptation strategy that provides most quality for the end user. The results are verified by visual inspection and by objective quality evaluation using SPSNR.




(a) (b)

**Fig. 3**. Rate-distortion diagrams for strategy selection. Anchor nodes are represented along with the corresponding cubic polynomial value functions. The client resources under analysis are highlighted using vertical lines. (a) *Hall monitor*. (b) *Highway*.

The test sequences are *Hall Monitor* from the MPEG–4 test sequences, and *Highway* from the MPEG–7 test sequences. Both sequences are in CIF format at 25 Hz; the length is 300 frames. In our experiments, the following frame-based adaptation strategies are compared: (a) coded original sequence; (b) spatial resolution reduction; (c) semantic prefiltering with lowpass-filtering; (d) semantic prefiltering with static background. A single resource, i.e. bitrate, is considered. In order to assess the performance of the selection mechanism both for low-quality and for high-quality video, the bitrate of the client has been set to $R^{\text{UMTS}}_{\text{client}} = 176$ Kbit/s and to $R^{\text{ADSL}}_{\text{client}} = 1000$ Kbit/s. The former corresponds to the bandwidth supported by the UMTS multimedia protocol. The latter is sometimes used for video streaming over asymmetric digital subscriber lines (ADSL).

The anchor nodes have been calculated for the following bitrates: 150, 200, 250, 300, 500 and 1000 Kbit/s. In the rate-distortion diagrams in Fig. 3, each data point represents one anchor node. A cubic function is further fit to the anchor nodes of each adaptation strategy.

For *Hall monitor*, evaluating the value function at $R_{\text{client}} \leqslant 176$ Kbit/s leads to the following maximal SPSNR: 27.4 dB for coded original (a); 22.6 dB for spatial resolution reduction (b); 25.3 dB for semantic prefiltering with lowpass-filtering (c); 29.4 dB for semantic prefiltering with static background (d). Thus, according to Eq. (2), the adaptation strategy that provides most perceived quality for the end user is semantic prefiltering with static background (d). At $R_{\text{client}} \leqslant 1000$ Kbit/s, the maximal SPSNR is: 35 dB for the coded original; 22.8 dB for spatial resolution reduction; 27.2 dB for semantic prefiltering with lowpass-filtering; 35.6 dB for semantic prefiltering with static back-

ground. Thus, the selected adaptation strategy is semantic prefiltering with static background (d) as well.

For *Highway*, the resource allocation problem is solved in a similar way. The adaptation strategies that provide most quality are found to be semantic prefiltering with static background (d) at 176 Kbit/s, and the coded original sequence (a) at 1000 Kbit/s. These results are next verified by visual inspection and by objective quality evaluation. The former is done by inspecting sample frames from sequences coded using different strategies. The latter is achieved by measuring SPSNR at 176 Kbit/s and at 1000 Kbit/s for each strategy. In Fig. 4, sample frames are shown for the sequence *Hall monitor*. At 176 Kbit/s (left column), the person's face and the monitor have slightly more details with the semantic strategies (c) and (d) than with the non-semantic strategies (a) and (b). Also, the background is severely corrupted by coding artifacts in the coded original (a). This is particularly visible on background edges. At 1000 Kbit/s (right column), spatial resolution reduction (b) and lowpass-filtered background (c) have substantially lower quality than the coded original (a) and static background (d). On the other hand, it is difficult to perceive differences between the coded original (a) and static background (d). In Fig. 5, sample frames are shown for the sequence *Highway*. At 176 Kbit/s, the background of semantic prefiltering with static background (d) has higher quality than the background of the coded original (a). In particular, the white painted lines on the road are sharper with static background (d). At 1000 Kbit/s however, the shadow cast by the truck stops in an unnatural way in static background (d). These artificial boundaries result from the object segmentation process used by the semantic prefiltering step. These boundaries are visually annoying and lead to a lower perceptual quality for static background (d) than for the coded original (a).

The SPSNR for the two test sequences coded at 176 Kbit/s and at 1000 Kbit/s using different adaptation strategies is given in Table 1. As expected, the highest objective quality for *Hall monitor* is achieved by using semantic prefiltering with static background at both 176 Kbit/s and 1000 Kbit/s. For *Highway*, the highest SPSNR obtained by using semantic prefiltering with static background at 176 Kbit/s, and by the coded original at 1000 Kbit/s.

| BITRATE | 176 Kbit/s | 1000 Kbit/s |
|---|---|---|
| **Hall monitor** | | |
| *Coded original* | 27.5 dB | 35.0 dB |
| *Spatial resolution reduction* | 22.7 dB | 22.8 dB |
| *Lowpass-filtered background* | 25.3 dB | 27.2 dB |
| *Static background* | 29.4 dB | 35.6 dB |
| **Highway** | | |
| *Coded original* | 29.0 dB | 35.8 dB |
| *Spatial resolution reduction* | 23.4 dB | 23.5 dB |
| *Lowpass-filtered background* | 27.7 dB | 30.1 dB |
| *Static background* | 29.8 dB | 35.1 dB |

**Table 1**. SPSNR for the sequences *Hall monitor* and *Highway* coded at 176 Kbit/s and at 1000 Kbit/s using different adaptation strategies.

## 4. CONCLUSIONS

We presented a video encoding scheme that uses object segmentation based on motion to increase the perceived quality of surveillance video as well as to meet the restricted capabilities of mobile devices. The scheme is used to select among different adaptation strategies in realistic content delivery situations and has been demonstrated on surveillance test sequences. Both visual inspection and objective quality evaluation results confirm that the adaptive delivery framework described in this paper is capable to determine the adaptation strategy that leads to the best perceptual video quality. In fact, the strategies that have been selected for delivery have also the highest SPSNR in all tested cases. As part of our future work, we would like to point out that the discussed method requires quality to be computed explicitly for each candidate strategy. Such calculations are time-consuming and need at the moment to be performed offline. A solution to this problem is *quality function prediction*, where the quality is estimated for each strategy based on content features instead of being actually computed.

## 5. REFERENCES


[1] Yao Wang, Jörn Ostermann, and Ya-Qin Zhang, *Video Processing and Communications*, Prentice Hall, Upper Saddle River, USA, 2001.

[2] Anthony Vetro, Charilaos Christopoulos, and Huifang Sun, "Video transcoding architectures and techniques: an overview," *IEEE Signal Processing Magazine*, vol. 20, no. 2, pp. 18–29, March 2003.

[3] Anthony Vetro, Huifang Sun, and Yao Wang, "Object-based transcoding for adaptable video content delivery," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 387–401, March 2001.

[4] R. Cucchiara, C. Grana, and A. Prati, "Semantic video transcoding using classes of relevance," *International Journal of Image and Graphics*, vol. 3, no. 1, pp. 145–169, January 2003.

[5] Andrea Cavallaro, Olivier Steiger, and Touradj Ebrahimi, "Semantic video analysis for adaptive content delivery and automatic description," *IEEE Transactions on Circuits and Systems for Video Technology*, 2005 (to appear).

[6] Andrea Cavallaro, Olivier Steiger, and Touradj Ebrahimi, "Tracking video objects in cluttered background," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 4, April 2005.

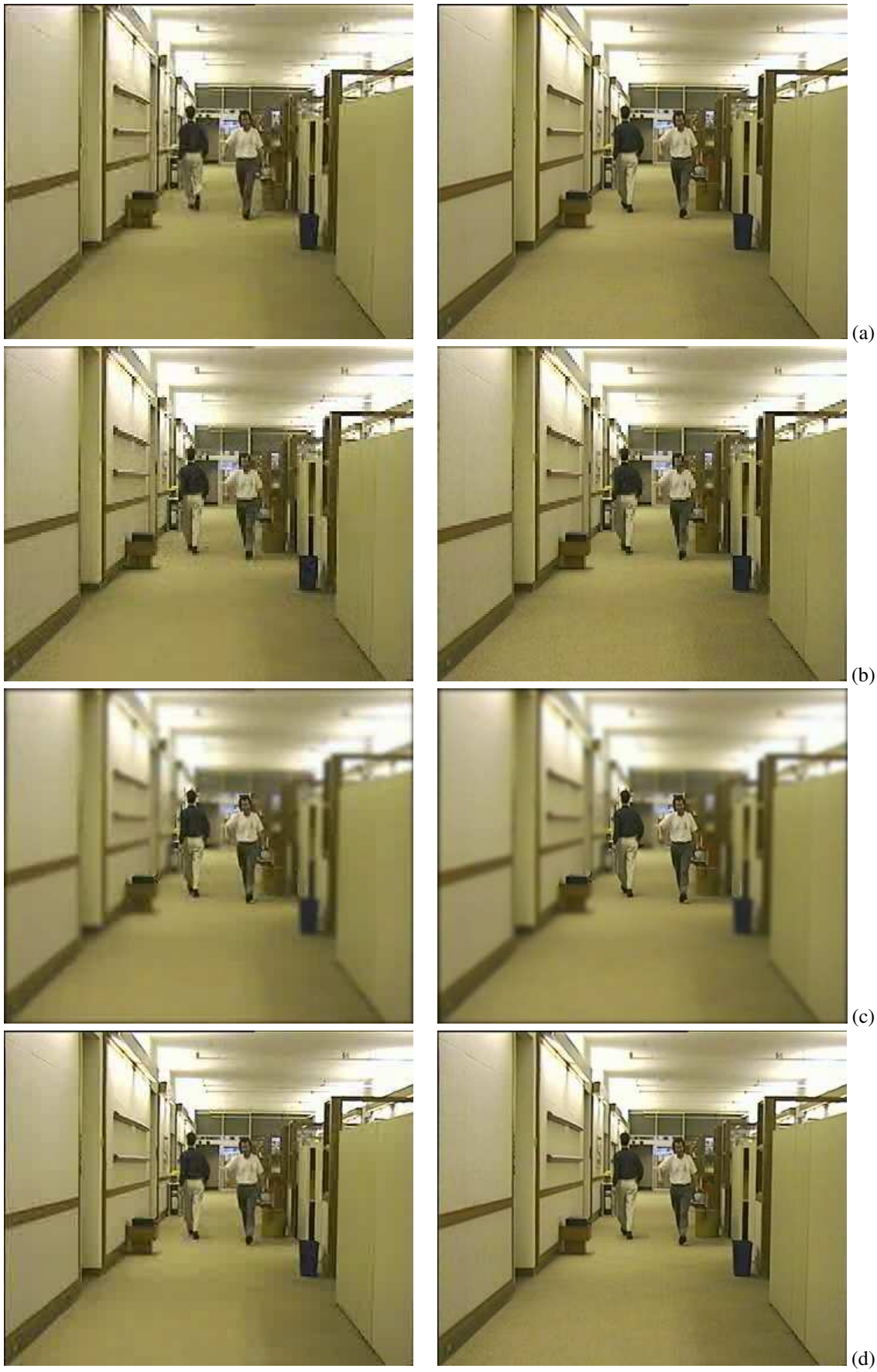



**Fig. 4**. Frame 190 from *Hall monitor* for different adaptation strategies. The coding bitrates are: (left column) 176 Kbit/s; (right column) 1000 Kbit/s. The strategies under analysis are: (a) Coded original. (b) Spatial resolution reduction. (c) Lowpass-filtered background. (d) Static background.

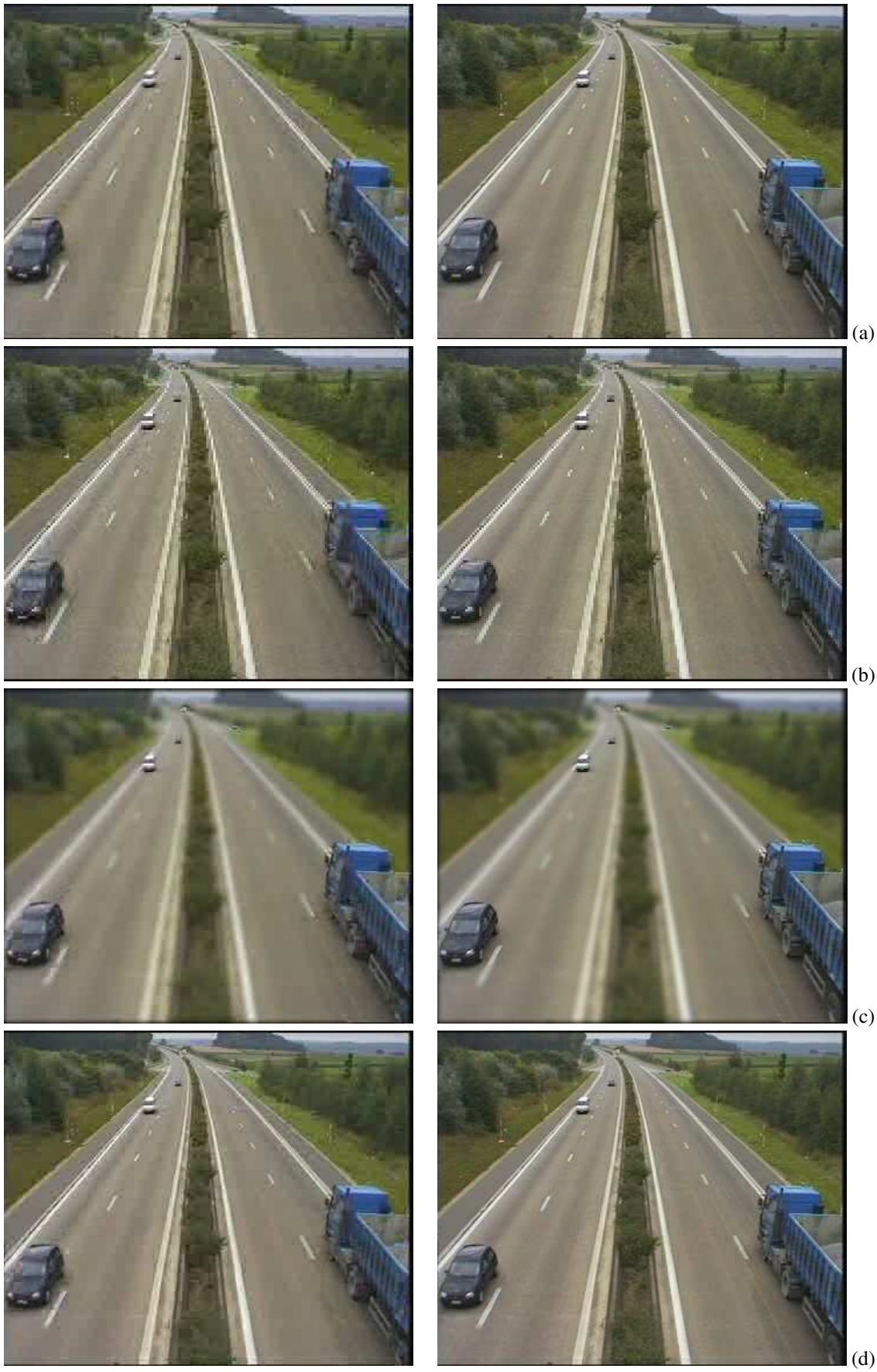



**Fig. 5**. Frame 20 from *Highway* for different adaptation strategies. The coding bitrates are: (left column) 176 Kbit/s; (right column) 1000 Kbit/s. The strategies under analysis are: (a) Coded original. (b) Spatial resolution reduction. (c) Lowpass-filtered background. (d) Static background.