TOWARDS REGION-BASED CODING FOR VERY LOW BIT RATE VIDEO SERVICES

TOWARDS REGION-BASED CODING FOR VERY LOW BIT RATE VIDEO SERVICES

Janusz Konrad, Abdol-Reza Mansouri and Eric Dubois

Université du Québec
Institut national de la recherche scientifique
INRS-Télécommunications
16 Place du Commerce, Verdun
Québec, Canada, H3E 1H6

Août 1995

Rapport technique de l'INRS-Télécommunications no. 95-13

This is a limited-circulation report submitted to Bell-Northern Research to describe work carried out between September 1994 and August 1995 under contract No. 94-07 *"Very Low Bit Rate Video Services - Phase II"*.

Summary

In the first phase of this project we have compared the performance of the H.261 ($p\times$ 64kb/s) low bit rate video coding standard with the so-called Simulation Model 3 (SIM-3) proposed within the COST-211 activity of the European Community. Both are derived from the well-known block-based DPCM/DCT hybrid coding scheme and differ primarily in the motion model used. In extensive experiments the SIM-3 coder has been judged to significantly outperform the H.261 coder in terms of image quality. The improvement was especially due to the reduced image "blockiness". However, at very low rates (i.e., below 16kb/s) the performance of the SIM-3 coder was still unsatisfactory for the critical parts of *"Carphone"* and *"Foreman"* sequences.

In the phase II of the project we have identified two goals: (1) to find the performance limits of the H.261 standard with respect to motion estimation, and (2) to explore the applicability of region-oriented schemes to the coding of videophone sequences at very low rates.

Thus, first we describe a block-based motion estimation that attempts to optimize the overall bit budget for intensity residual (prediction error), motion vectors and overhead information. We compare simulation results for this scheme with full-search block matching in the context of H.261 coding. We conclude that a gain of at most 1dB is possible and this at higher bit rates (32-64kb/s); at very low rates no clear improvement was observed. Since this result has been obtained by exhaustive search, any practical scheme would result in a smaller gain.

Then, we move to region-based representations of video sequences that are expected to eliminate image blockiness at very low rates. Such representations are at the center of interest of the MPEG-4 forum established to study methods for very low bit rate audio-visual coding. In this project we have not attempted to simulate a complete region-based video coder, but rather study some of its essential blocks. We describe a region-based representation of motion and propose a new algorithm for the computation of such representation's parameters. Its essence is based on the observation that usually motion boundaries correspond to intensity boundaries. Therefore, we describe an intensity-based image segmentation and we give details of a three-step method for the computation of motion parameters for each region of the segmentation. Since a complete region-based motion representation requires also the description of region shape, we compare several schemes (lossless and lossy) for efficient shape representation. We conclude that region-based motion estimation provides a substantial reduction of the prediction error, however at the same time we demonstrate that a large portion of the bit rate has to be allocated to the representation of region shape. It is not clear at this time whether the loss can be compensated by the achieved gain.

In the final sections of the report we describe visual demonstrations of some results, we discuss the results obtained and we propose directions for further research in the field of very low bit rate video services.

Contents

1 Introduction **1**

2 Rate-based motion vector estimation **3**

3 Region-based representation of video sequences **5**
3.1 Intra-frame image segmentation 5
3.2 Region-based motion representation and estimation 8
3.2.1 Estimation of initial motion parameters 8
3.2.2 Region fusion under motion constraint 9
3.2.3 Region boundary adjustment 10
3.2.4 Results 11
3.3 Efficient representation of region boundaries 15
3.3.1 Lossless coding 15
3.3.2 Lossy coding 15

4 Results for complete video sequences **17**
4.1 Region-based motion estimation and prediction 17
4.2 Representation of region-boundaries 20

5 Demonstration of results **23**

6 Discussion **26**

7 Further directions **27**

List of Figures

1 Block diagram of a possible region-based video coding scheme. Highlighted blocks have been studied in this phase of the project. 2

2 One image from each of the three QCIF sequences used in simulations. 4

3 Comparison of H.261 coder performance for standard SSD-based (sum of squared differences) and rate-based (RB) motion estimation: (a) scatter plot and local average fit (Lowess smoothing) for SSD-based motion estimation and *"Miss America"*; and the same fit for both algorithms for (b) *"Miss America"* and (c) *"Carphone"*. 4

4 Coarse, medium and fine intra-frame (intensity-based) segmentation maps for images from Fig. 2 computed by the MDL algorithm. 9

5 Two consecutive images from *"Miss America"*, their difference and the initial, after fusion and final segmentations, motion fields, prediction and error images. 12

6 Two consecutive images from *"Carphone"*, their difference and the initial, after fusion and final segmentations, motion fields, prediction and error images. 13

7 Two consecutive images from *"Foreman"*, their difference and the initial, after fusion and final segmentations, motion fields, prediction and error images. 14

8 Intra-frame (intensity-based) segmentation maps for image #0 from the three test sequences. 16

9 Original image #90 from *"Miss America"*, its intensity-based segmentation map, corresponding contours and approximations for γ=7, 10 and 13. 18

10 Peak prediction gain of the motion-compensated prediction error for block-based, pixel-based and region-based (lossless and lossy) motion estimates for the three test sequences. 19

11 Original image # 171 from *"Carphone"*, initial and final segmentation maps, motion fields, motion-compensated predicted images and prediction errors. 21

12 Motion-based (final) segmentation maps from Figs. 5, 6 and 7 after lossy compression. Average bit-rates are given in Table 8. 25

List of Tables

1 Bit rates per frame for intra-frame segmentation maps from Fig. 8 encoded using pulse-code modulation (PCM), run-length (RLC), simplified binary-arithmetic (BAC) and chain (CC) coding. To encode all the labels from Fig. 8 using the PCM, 5 bits have been used for *"Miss America"* and 7 for the other sequences. 16

2 Bit rates per frame (bits/fr) for intra-frame segmentation maps from Fig. 8 encoded using lossy compression for three values of γ. δ is a distortion measure. 17

3 Average number of regions and contour points for intra-frame segmentations of the test sequences. 20

4 Average rates over sequences of intra-frame segmentations encoded using the three lossless methods. *Starting points of chains are not accounted for. The entropy is an estimate of the entropy rate under the assumption that the process consisting of direction chain codes is stationary and first-order Markov. 22

5 Average rates over sequences of intra-frame segmentations encoded using the lossy method for three values of γ. 22

6 Average number of regions and contour points for motion-based segmentations of the test sequences. 23

7 Average rates over sequences of motion-based segmentations encoded using the three lossless methods. *Starting points of chains are not accounted for. The entropy is an estimate of the entropy rate under the assumption that the process is stationary and first-order Markov. 24

8 Average rates over sequences of motion-based segmentations encoded using the lossy method for three values of γ. 24

1 Introduction

In the last few years novel and very demanding applications, such as desktop video-conferencing and mobile communications, have given a renewed impetus to the development of very low bit rate video coding algorithms. The bit budget is very restricted in such applications and therefore every transmitted bit carries a significant amount of information. Due to the nature of contemporary coders this information is related to motion parameters, to intensity (and color) residual or to overhead information. In this phase of the project (phase II) we concentrate on the motion aspect and we explore issues related to the modeling and computation of motion for very low bit rate applications.

From the phase I of this project [16] we have learned that the motion model used has a decisive impact on the quality of the encoded video sequences at low bit rates. The above observation confirms one of the current goals in very low bit rate video coding research: to establish improved and flexible motion models. In the near term, algorithms working within the H.261 [1] standard or around it, such as the H.263 activity, are being developed. These new algorithms usually require modest decoder modification, and therefore are rather inexpensive in practical implementation. Interesting examples of schemes requiring such a modification are described in the literature [19, 18]. In this context we discuss in Section 2 our preliminary study of the performance limits of the H.261 standard with respect to motion estimation. We compare the rate-distortion performance of the H.261 coder for two motion estimation algorithms: the standard exhaustive search (with respect to the sum of squared differences) and a rate-based algorithm, that minimizes the total number of bits needed for motion, intensity residual and overhead information per block.

As for the long-term goal, models and algorithms very different from the H.261 standard are investigated in search of a performance improvement. This search is motivated by the inherent limitations of standard block-based motion-compensated coding schemes: the crude "object" model, which assumes that image is made up of regularly-sized moving square patches, and the simple motion model, which assumes a translational displacement. These two models result in very annoying "blockiness" of image intensities at very low bit rates; the human visual system (HVS) is very sensitive to structured patterns such as the "staircase" effect. To overcome these limitations, object- and region-based methods have been proposed in the literature [17, 6, 21]. In such methods, motion of 3-D objects (projected onto 2-D image plane) or motion of 2-D regions (in the image plane) is modeled, computed and used for prediction or tracking. We consider the region-based approach not only as the means to reduce "blockiness" of image intensities, but also as a way of hiding coding artifacts. More precisely, we believe that due to the spatial masking property of the HVS the coding errors, most prominent around sharp intensity transitions (region boundaries), should be less visible. Therefore, in Section 3 we discuss our approach to motion estimation and segmentation that is adapted to regions; the moving "objects" are regions related to intensity segmentation and the motion model is an affine transformation on the

plane. The proposed method is based on intra-frame intensity segmentation followed by region-based motion estimation, motion-based fusion and boundary adjustment. The resulting region-specific motion assures a substantial improvement in prediction quality, but simultaneously raises two important issues: coding of image intensities supported on each region and coding of image segmentation maps. The former issue has been addressed in a number of publications over the last few years [9, 2, 4, 22], although it has not yet found a satisfactory solution for practical applications. As for the coding of segmentation maps, in Section 3.3 we review lossless techniques based on run-length, simplified arithmetic and chain coding, as well as a lossy method based on polygonal region approximation.

Fig. 1 shows the block diagram of a region-based video coding scheme that we are exploring in this project. As usual it comprises two compression branches: intra- and inter-frame. Both require a segmentation procedure to extract intensity- or motion-based regions, and a coding scheme for efficient transmission of segmentation maps. Additionally, motion estimation and coding of motion parameters must be performed in the inter-frame branch. Suitable coding of intensity and color must be assured in both branches. The blocks studied in detail within the mandate of this project are highlighted.

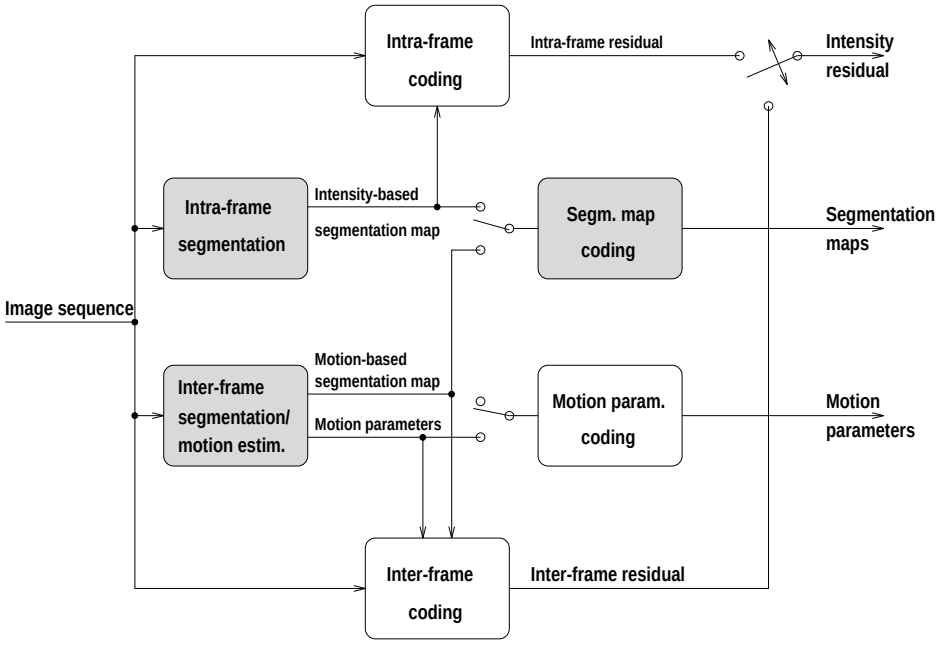


Figure 1: Block diagram of a possible region-based video coding scheme. Highlighted blocks have been studied in this phase of the project.

2 Rate-based motion vector estimation

In most implementations of motion-compensated video coders, the motion vectors are selected to minimize a measure of prediction error, for example the sum of the absolute values of the prediction error over a block. While this is reasonably effective in minimizing the number of bits required to represent the motion-compensated prediction error (or residual), it is not necessarily optimal when both the bits to represent the residual and to represent the motion vectors are considered. If several motion vectors give very nearly the same residual, the one which requires the fewest number of bits to represent should be selected. Of course, this depends on how motion vectors are encoded, and can only have an impact if a variable-length coding scheme is used.

The above problem is very general and is currently the subject of active research [7]. The goal in this project is to carry out a preliminary investigation to see how this approach can be applied in the context of standard low bit rate coding schemes such as H.261 and the forthcoming H.263, and what impact it has on rate-distortion performance of the coder.

The approach we have taken is straightforward but computationally intensive. The goal of the experiment is to determine what can be gained before trying to develop low-complexity implementations. Within the H.261 coding syntax, we compare two different motion estimation algorithms:

- exhaustive search for the minimum of the sum of squared differences (SSD) of the prediction error over a macroblock,
- exhaustive search for the minimum of rate $R = R_r + R_m + R_o$ for a macroblock, where R_r is the number of bits required to code the intensity/color residual (in case the macroblock is inter-coded), R_m is the number of bits required to code the differential motion vectors (in case motion compensation is used), and R_o is the control information that is packaged into the macroblock header.

Thus, in the rate-based (RB) scheme, for each candidate motion vector the number of bits required to represent it is computed, along with the number of bits required to code the resulting residual using the standard DCT method with a fixed quantization parameter and all other overhead bits; the motion vector giving the lowest total number of bits is selected for the macroblock. In other words, whereas in typical H.261 motion estimation is performed so as to minimize an SSD measure of prediction error between macroblocks, we are now interested in estimating the motion vector which yields the shortest code for the macroblock to be coded.

Implementing the above rate-constrained scheme entails exhaustively precoding the macroblock to be coded, with each possible choice of motion vector. Whereas in standard H.261 a simple SSD has to be computed for each choice of motion vector, we now have to compute the Discrete Cosine Transform (DCT) of the macroblock, quantize the resulting DCT coefficients and Variable-Length Code (VLC) them. This is done approximately 30x30 = 900 times for each macroblock.

Since H.261 uses a predictive coding of motion vectors for each macroblock, with shorter codewords for smaller differences, this approach would favor, among several

motion vectors giving similar residual energy, the one most similar to the horizontally preceding motion vector. Note that although the quantization parameter is fixed throughout, the actual distortion may vary depending on the details of the residual. Ideally, one should also vary the quantization parameter to keep the distortion for the macroblock constant as we compare different motion vectors, but we did not try this due to the great increase of computations that it would require.

"Miss America", image #6 "Carphone", image #3 "Foreman", image #3

Figure 2: One image from each of the three QCIF sequences used in simulations.

Figure 3: Comparison of H.261 coder performance for standard SSD-based (sum of squared differences) and rate-based (RB) motion estimation: (a) scatter plot and local average fit (Lowess smoothing) for SSD-based motion estimation and *"Miss America"*; and the same fit for both algorithms for (b) *"Miss America"* and (c) *"Carphone"*.

The experiment was carried out on two sequences (50 frames of *"Miss America"* and 100 frames of *"Carphone"* at 3:1 temporal subsampling, Fig. 2) with 6 different quantizer parameters. Since both rate and distortion are different between the rate-based and the standard algorithm, we have compared the average rate-distortion performance of the two methods. Fig. 3(a) shows the scatter plot of rates and distortions obtained over the 49 coded P frames of *"Miss America"*, along with a smoothed version (Lowess smoothing). Figures 3(b) and 3(c) show the smooth rate-distortion

curves for the standard full-search motion estimation (SSD) and the rate-based motion estimation (RB) for *"Miss America"* and *"Carphone"*, respectively. As can be seen from this figure the performance improvement is at higher quality; at a fixed rate there is about 1.3dB $PSNR$ gain for *"Miss America"* and 0.8dB for *"Carphone"* while at a fixed $PSNR$ the reduction of bit rate is about 1kb and 2kb, respectively. At lower quality, however, there is no appreciable difference between the two algorithms which may be partially due to the role that the overhead information plays at the lower rates. Since in our experiments a very time-consuming exhaustive search was used, any practical scheme approximating the above algorithm would attain a smaller gain.

3 Region-based representation of video sequences

In traditional coding schemes, the most annoying artifact at very low rates is the "blockiness" of image intensities. The visibility of blocks is particularly objectionable at object boundaries since the HVS is sensitive to regular patterns such as the "staircase" effect. Therefore, it is essential that object boundaries be represented more precisely. At the same time, a certain level of distortion may be permitted around the boundary due to the spatial masking of the HVS. Below, we propose an approach to region-based representation of motion. It is based on the observation that only under specific circumstances motion boundaries will not coincide with intensity edges [10]. Thus, we study three blocks from Fig. 1 that are pertinent to the region-based representation of motion:

- intensity-based (intra-frame) segmentation,
- inter-frame segmentation and motion estimation,
- coding of segmentation maps.

To compute the intensity-based segmentation we implement a novel algorithm recently proposed in the literature (Section 3.1). We use the resulting segmentation as an initial step for our new 3-step algorithm: motion estimation for intensity-derived regions, motion-based region fusion and adjustment of region boundaries (Section 3.2). Since the estimated motion parameters are region-specific, image partition into regions must be transmitted as well. We compare four techniques used for compression of such partitions: three lossy and one lossless (Section 3.3).

3.1 Intra-frame image segmentation

A very popular technique for image segmentation is called split-and-merge. It is based on partitioning the image into square blocks and further splitting or merging these blocks based on some homogeneity criterion. This criterion is often chosen to be the variance of the intensity within the block. If the variance in the block is larger than a threshold, the block is split into four sub-blocks, to which the splitting criterion is recursively applied. If on the other hand two neighboring blocks can be merged

while keeping the overall variance within some bounds, then they are merged. The split and merge procedures are iteratively applied until no further splits or merges are possible. Since our goal is to search for an image segmentation that is not block-based, we would need to carry out block splitting down to the pixel level. At such a level, however, the method does not perform well especially for complex images with significant noise content.

Therefore, we use a different approach that still exploits pixel interdependency but at a lower level. We partition each image in a video sequence into regions of almost uniform intensity by computing a Minimum Description-Length (MDL) estimate of the image [15]. Let $g(\boldsymbol{x}, t)$ be image intensity at spatial position $\boldsymbol{x}$ and time t. Defining the description $\mathcal{S}$ to be a piecewise-constant image function in $\boldsymbol{x}$, and assuming the image g to be of the form $g(\boldsymbol{x}, t) = \mathcal{S}(\boldsymbol{x}, t) + n(\boldsymbol{x}, t)$ with n a stationary zero-mean Gaussian white noise process of variance σ^2, the conditional description length $L(\mathcal{S}/g)$ is given by

$$L(\mathcal{S}/g) = L(g/\mathcal{S}) + L(\mathcal{S}) = \sum_{i=1}^{P} [g(\boldsymbol{x}_i, t) - \mathcal{S}(\boldsymbol{x}_i, t)]^2/(2\sigma^2) + \frac{1}{2} log(2\pi\sigma^2) + L(\mathcal{S}), \quad (1)$$

corresponding to entropy coding the (independent) random variables $n(\boldsymbol{x}_i, t)$ with P being the number of pixels considered. The description length $L(\mathcal{S})$ of the image g is cast in algorithmic, as opposed to statistical, terms, and is defined as a measure of complexity. As such a measure we adopt here [15]:

$$L(\mathcal{S}) = \frac{1}{2} \sum_{i=1}^{P} \sum_{\boldsymbol{x}_j \in \eta_2(\boldsymbol{x}_i)} (1 - \delta(\mathcal{S}(\boldsymbol{x}_i, t) - \mathcal{S}(\boldsymbol{x}_j, t))),$$

where $\eta_2(\boldsymbol{x}_i)$ is a second-order (8-connected) neighborhood of $\boldsymbol{x}_i$ and $\delta(x) = 1$ for $x = 0$ and 0 otherwise. The above measure is proportional to the number of boundary points and reflects well the complexity of boundaries. In order to minimize the total description length $L(\mathcal{S}/g)$, the δ function in $L(\mathcal{S})$ is regularized through a parameterized family e^{-x^2/ϵ^2} of Gaussians approaching δ as $\epsilon \to 0$ [15]. The MDL estimate of the image is computed by successive continuation on ϵ together with a Gauss-Seidel iterative descent algorithm, as indicated in [15]. In order to compute an estimate close to that which minimizes (1), continuation has to be done very slowly. This implies that a large number of (small) continuation steps have to be provided, so as to allow the approximating Gaussians to approach the 0-set indicator function. We accelerate the computation of $\mathcal{S}$, by stopping the continuation on ϵ after a certain number of iterations, and following with unsupervised intensity clustering. Intensity clustering can itself be performed according to different algorithms. We choose one which has an intuitive explanation as minimizing some energy function. We define the energy function of interest as:

$$U(V, \mu/\mathcal{S}) = \sum_{i=1}^{P} \sum_{m=1}^{M} V_m(\boldsymbol{x}_i, t)[\mathcal{S}(\boldsymbol{x}_i, t) - \mu(m)]^2, \quad (2)$$

where M is the maximum number of regions, $\mu(m)$ are the centers of intensity clusters and $V_m(\boldsymbol{x}_i, t)$ are binary variables indicating whether location $(\boldsymbol{x}_i, t)$ is assigned to cluster m. The variables $V_m(\boldsymbol{x}_i, t)$ are constrained in that each location $(\boldsymbol{x}_i, t)$ must be assigned to exactly one cluster. By computing the Gibbs distribution corresponding to the above energy function, we perform joint estimation of the binary variables $V_m(\boldsymbol{x}_i, t)$ and the continuous variables $\mu(m)$ via the EM (expectation-maximization) algorithm [5]. The Gibbs distribution corresponding to U is given by:

$$p(V, \mu / \mathcal{S}) = \frac{1}{Z} e^{-\beta U(V, \mu / \mathcal{S})},$$

where β is the inverse-temperature and Z is the partition function. The EM algorithm is used to estimate parameters when all of the data is not available. It yields a sequence of estimates which converge to a local maximum likelihood estimate. In the context of clustering, the problem is that the cluster centers μ have to be found at the same time as the cluster elements themselves (V). If either the binary variables V or the continuous variables μ were known, the energy function U above could have been minimized with respect to the remaining set of variables. The EM update equations are as follows:

$$\begin{aligned} E(V_{im}^k) &= \frac{e^{-\beta(\mathcal{S}(\boldsymbol{x}_i,t) - \mu^{k-1}(m))^2}}{\sum_l e^{-\beta(\mathcal{S}(\boldsymbol{x}_i,t) - \mu^{k-1}(l))^2}} \\ \mu^k(m) &= \frac{\sum_i E(V_{im}^k) \mathcal{S}(\boldsymbol{x}_i, t)}{\sum_i E(V_{im}^k)} \end{aligned}$$

The first equation above corresponds to computing the mean of the binary variable V_{im}^k, while the second equation corresponds to computing a weighted average of the $\mathcal{S}(\boldsymbol{x}_i, t)$. These two computations are iterated, one after the other, until convergence is attained (typically after 20 or so iterations). In the equations above, the superscript k indicates the iteration number. At the same time, we perform continuation on the inverse-temperature parameter β. We start with a small value of β, compute the EM estimates, increase the value of β, compute the new EM estimates using the previous ones as starting points, increase β again, and so on. We usually perform continuation in 100 steps, although experimental results show that fewer steps would suffice. Note that other algorithms, such as K-means clustering, may be used to minimize the above energy function (2). However, K-means clustering is nothing but the zero-temperature version of the above algorithm.

The result of clustering, namely the binary variables $V_m(\boldsymbol{x}_i, t)$ describe image partition Ψ into N connected regions as follows: each region $\psi_n \in \Psi$ $(i = 1, ..., N)$ consists of all sites $(\boldsymbol{x}_i, t)$ such that $V_n(\boldsymbol{x}_i, t)$=1. From the partition Ψ a piecewise-constant description $\mathcal{S}$ can be recovered: $\mathcal{S}(\boldsymbol{x}_i, t) = n$ if $\boldsymbol{x}_i \in \psi_n$.

Following intensity clustering, the regions in the piecewise-constant description (segmentation) $\mathcal{S}$ are labeled. This labeling is done by assigning initial labels to points in the image, and by propagating these labels through neighborhoods. The whole process is accelerated by making use of different scan directions for label propagation.

In the remainder of this document the segmentation $\mathcal{S}$ and the partition Ψ will be used interchangeably.

Fig. 4 shows examples of intra-frame MDL-based segmentation for the three images from Fig. 2 and three different degrees of coarseness. The coarseness of segmentations depends on the variance σ^2 in the model (1); the higher the σ^2, the coarser the segmentation. Note a good correspondence between the computed regions at medium coarseness and the human perception of regions in those images (fig. 2). The fine segmentation may be judged too fine (too many regions) for some applications. Here, however, we assume that a motion boundary coincides with an intensity boundary and consequently that an image oversegmented with respect to intensity incorporates all motion boundaries. Since this is later corrected by motion-based fusion (Section 3.2.2), also oversegmented images can be used as the input to the motion estimation algorithm described below.

In the case of intra-frame coding (Fig. 1), however, oversegmentation may be undesirable. Then, the variance σ^2 needs to be chosen judiciously to assure sufficiently coarse segmentation. It is not clear what an optimal value of σ^2 should be; ideally one would like to jointly optimize segmentation maps and motion parameters to achieve the minimal overall bit rate. This is a difficult problem and is currently a topic of intense research.

3.2 Region-based motion representation and estimation

The region-based motion representation and estimation developed below are based on two underlying assumptions. First, we assume that in an image sequence motion boundaries almost always coincide with intensity boundaries. This observation is not new and has been exploited before in the computation of motion [10, 13]. Secondly, we assume that the displacement of each point $\boldsymbol{x} = [x, y]^T$ in an image region can be described by the following affine transformation [6]

$$\boldsymbol{d}(\boldsymbol{x}, \phi) = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} x - x_g \\ y - y_g \end{bmatrix}$$

where x_g, y_g are the coordinates of the center of gravity of the region. Thus, motion of each region ψ_n is described by the vector of parameters $\phi_n = [a_1, a_2, b_{11}, b_{12}, b_{21}, b_{22}]^T$. Other motion models could be incorporated into the algorithm presented below as well.

3.2.1 Estimation of initial motion parameters

Given the initial intensity-based partition Ψ at time t, we estimate motion parameters $\phi_1, \ldots, \phi_N$ that define the transformation of each region $\psi_n \in \Psi$ $(n = 1, \ldots, N)$ onto image frame at time t_-. Let the vector $\Phi = [\phi_1^T, ..., \phi_N^T]^T$ describe motion parameters of all regions in partition Ψ, i.e., ϕ_n describes motion of region ψ_n. Then, the estimation of parameters ϕ_n for each region in image g at time t can be carried

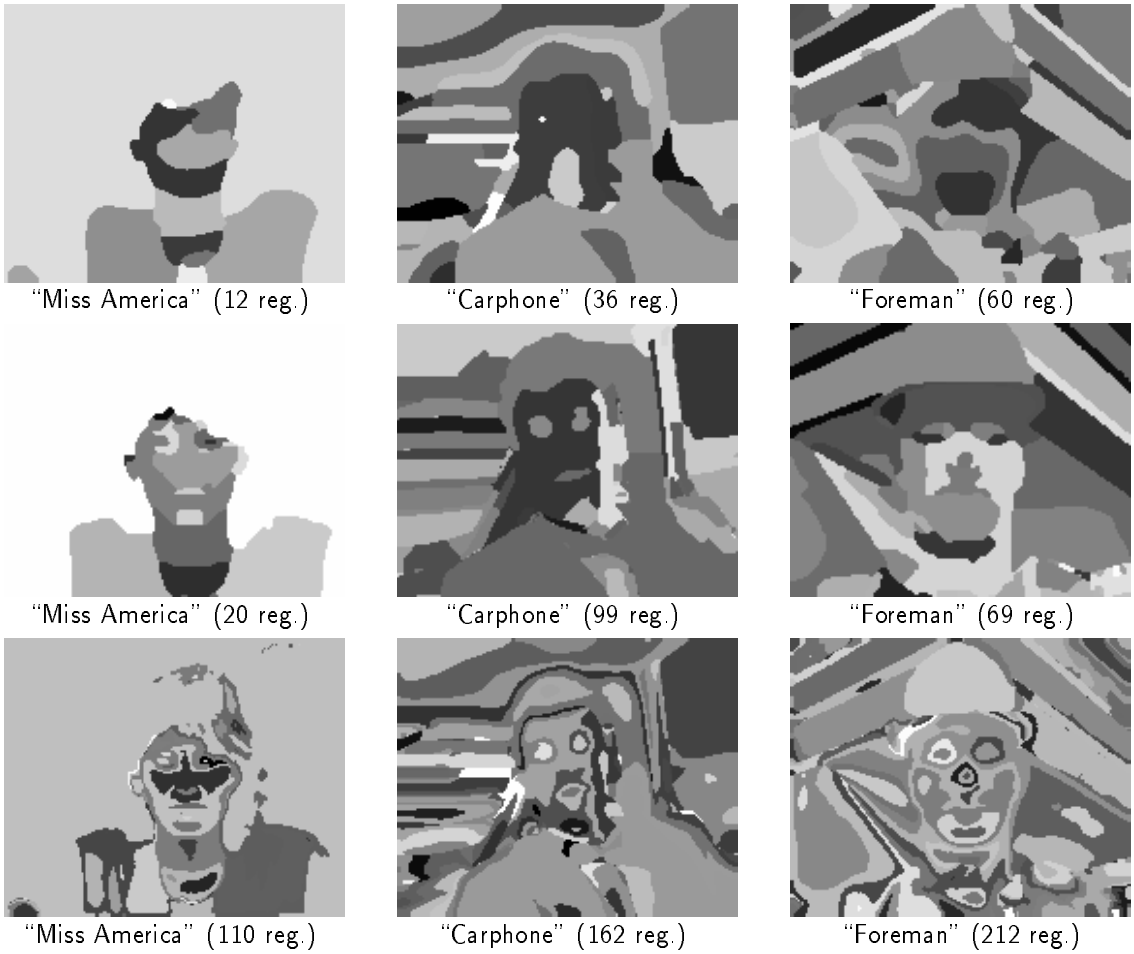


Figure 4: Coarse, medium and fine intra-frame (intensity-based) segmentation maps for images from Fig. 2 computed by the MDL algorithm.

out by minimizing the displaced pixel difference (DPD) for each region separately

$$\min_{\{\phi_n\}} \sum_{\boldsymbol{x}\in\psi_n} [\tilde{g}(\boldsymbol{x} - \boldsymbol{d}(\boldsymbol{x},\phi_n), t_-) - g(\boldsymbol{x},t)]^2, \quad \forall n, \tag{3}$$

where $\tilde{g}(\boldsymbol{x} - \boldsymbol{d}(\boldsymbol{x},\phi_n), t_-)$ denotes a pixel in image g at time t_- compensated for motion between t_- and t and spatially ($\tilde{g}$) interpolated by a C^1 bicubic operator [11]. The formulation (3) is similar to the one from [6], but here the minimization is carried out using the Gauss-Newton algorithm.

3.2.2 Region fusion under motion constraint

Since not all intensity boundaries correspond to motion boundaries, an elimination of redundant intensity boundaries needs to be performed. This corresponds to region fusion and is expected to improve precision of motion estimates; larger regions undergoing coherent motion provide more data to minimization in (3). At the same time, region fusion is expected to simplify the image partition Ψ. This is important

for coding since fewer region boundaries allows a smaller bit rate to be allocated to the transmission of partition Ψ.

In order that the fusion be effective, too many rather than too few intensity boundaries (regions) should be detected during the initial segmentation. This can be controlled by the parameter σ^2. A segmentation which is too coarse may result in one region encompassing several moving objects, a situation that cannot be subsequently corrected unless region splitting is allowed (like in split-and-merge). It is not clear, however, how to split a region into arbitrarily-shaped sub-regions.

The initial intensity-based partition Ψ and the associated motion parameters Φ are modified by fusing neighboring regions if such a fusion decreases the following cost function:

$$\min_{\{N,\Phi\}} \sum_{n=1}^{N} \sum_{\boldsymbol{x}\in\psi_n} [\tilde{g}(\boldsymbol{x}-\boldsymbol{d}(\boldsymbol{x},\phi_n),t_-)-g(\boldsymbol{x},t)]^2 + \lambda_1 N. \tag{4}$$

The second term on the right-hand side encourages the fusion of neighboring regions with similar motion parameters; in the limit, as $\lambda_1 \to \infty$, all the regions are fused together, while for λ_1=0, two regions are fused only if the DPD of the resulting region (with new motion parameters) is smaller than the sum of the two previous DPDs. The minimization (4) is carried out pairwise for all adjacent regions.

3.2.3 Region boundary adjustment

The simplified partition Ψ after fusion gives rise to more precise motion parameters; only a few regions with distinctive motion remain and the number of pixels per region is increased. We use the increased precision of motion estimates to adjust the intensity-derived boundaries and make them consistent with the motion of every region. The adjustments are not expected to be very large; at most a few pixels to either side. The adjustments are carried out locally at the boundary of each region by the following minimization:

$$\begin{aligned} \min_{k\in\Theta(\boldsymbol{x})} \quad & [\tilde{g}(\boldsymbol{x}-\boldsymbol{d}(\boldsymbol{x},\phi_k),t_-)-g(\boldsymbol{x},t)]^2 + \\ & \lambda_2 \sum_{\boldsymbol{y}\in\eta_2(\boldsymbol{x})} 1-\delta(k-\mathcal{S}(\boldsymbol{y},t)), \quad \forall \boldsymbol{x}\in\mathcal{B}(\psi_n),\forall n \end{aligned} \tag{5}$$

where $\Theta(\boldsymbol{x}) = \{\mathcal{S}(\boldsymbol{x},t)\} \cup \{\mathcal{S}(\boldsymbol{y},t), \boldsymbol{y}\in\eta_1(\boldsymbol{x})\}$ is the set of labels encountered at $\boldsymbol{x}$ and in the first-order neighborhood η_1 of $\boldsymbol{x}$, and $\mathcal{B}(\psi_n)$ denotes the boundary of region ψ_n. The second term describes the complexity of region boundaries and is inspired by Markov random field models used in motion estimation [12]. Minimization (5) is performed using Jacobi relaxation (the new estimates are not used until the whole field is computed) and exhaustive search over $\Theta(\boldsymbol{x})$ for each $\boldsymbol{x}$.

In other words, every (segmentation) pixel on the boundary $\mathcal{B}$ of region ψ_n is examined with respect to the sum of two costs: squared prediction error and boundary complexity. The state space Θ for each pixel consists of its own label and of the

labels assigned to its 8-connected neighbors. The motion parameters ϕ_n are not updated until the whole boundary is completed. The immediately preceding label modifications do not impact the current minimization since Jacobi relaxation is used.

3.2.4 Results

Figs. 5, 6 and 7 show results for each of the sequences tested. Motion estimation and motion-compensated prediction have been carried out between the reference frame (shown in Fig. 2) and the preceding frame. Each set of results comprises the following images:

- the preceding and the reference image between which motion is estimated, plus the difference between them (non-compensated for motion),
- three segmentations: initial (intra-frame), after fusion and final after boundary adjustment (motion-based),
- three motion fields computed from parameter vector Φ: initial, after fusion and final (all subsampled by 4),
- three (motion-compensated) prediction images: initial, after fusion and final,
- three (motion-compensated) prediction error images: initial, after fusion and final (all magnified by 2).

Note a dramatic reduction in the number of regions in the segmentation after fusion and a simplified shape of boundaries after boundary adjustment.

The increased size of regions (due to fusion) has an impact on the precision of motion fields. The initial estimation of Φ (minimization (3)) has detected small amount of motion in *"Miss America"*, modest body and background (in the window) motion in *"Carphone"* and relatively rapid global motion in *"Foreman"*. After region fusion (minimization (4)) and re-estimation of motion parameters (minimization (3)) there is a clear articulation of head movement in *"Miss America"* (although the number of outliers increases), improved consistency of background and body motion in *"Carphone"* and *"Foreman"*. After boundary adjustment and another re-estimation of Φ (minimizations (5) and (3)) a more uniform vector field is obtained for *"Miss America"*, however the change is modest. A dramatic change can be observed for the other two sequences. In *"Carphone"* head motion has become uniformly translational and the underestimated background motion in the window has been corrected. In *"Foreman"* the final motion field is almost uniformly translational; this global motion corresponds very closely to the camera pan present at the beginning of that sequence.

The subjective improvement of motion fields translates directly into the quality improvement of predicted images. Note in the initial predicted images and prediction errors the opened right eye in *"Miss America"*, spurious tree in the window of *"Carphone"* and spurious diagonal line (top-left corner), distorted right bark and eye in *"Foreman"*. The fusion process has partially corrected the eye opening in *"Miss America"* and has corrected the distortions in *"Foreman"*. Only after boundary adjustment and re-estimation of Φ have all those distortions disappeared.

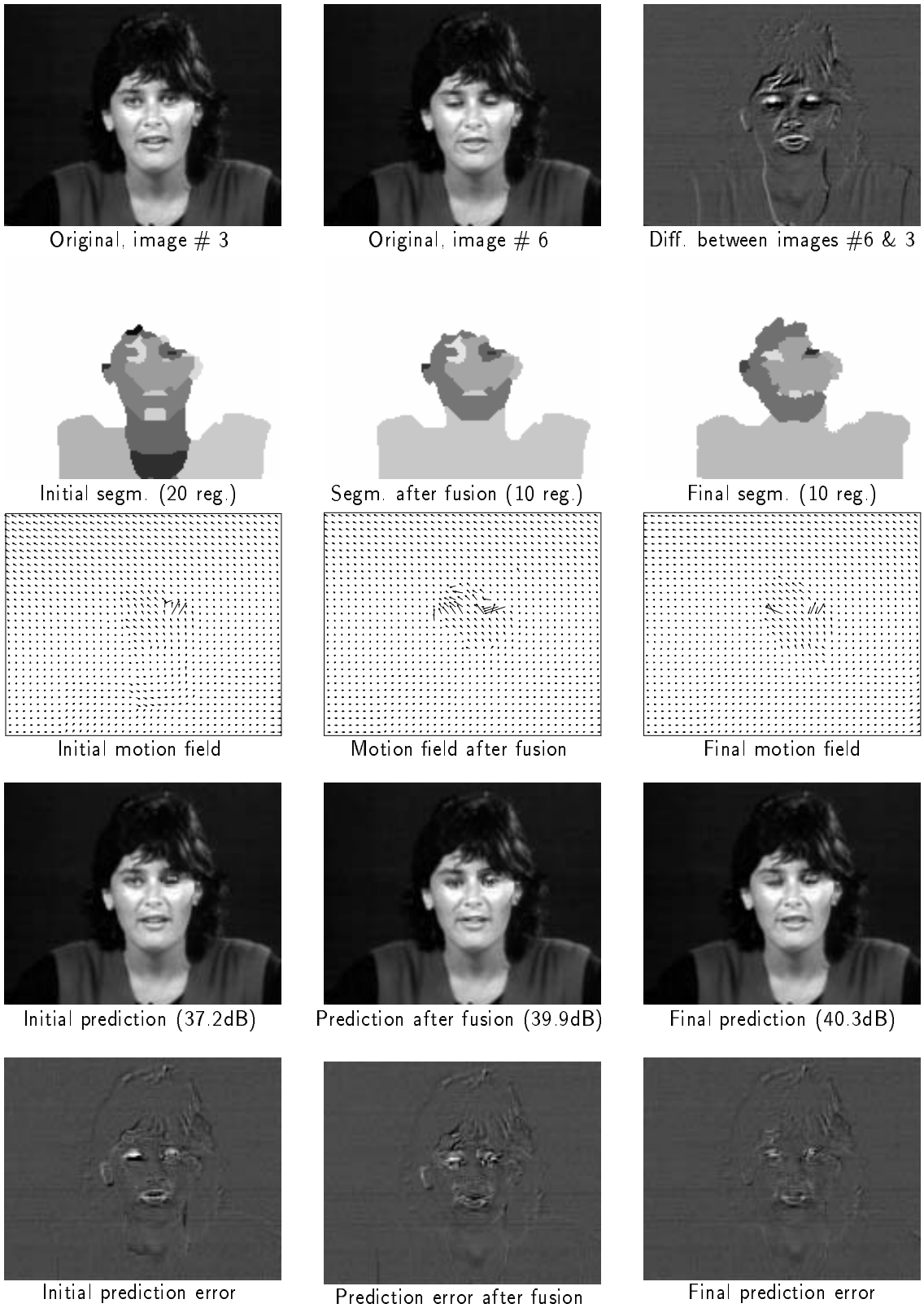


Figure 5: Two consecutive images from *"Miss America"*, their difference and the initial, after fusion and final segmentations, motion fields, prediction and error images.

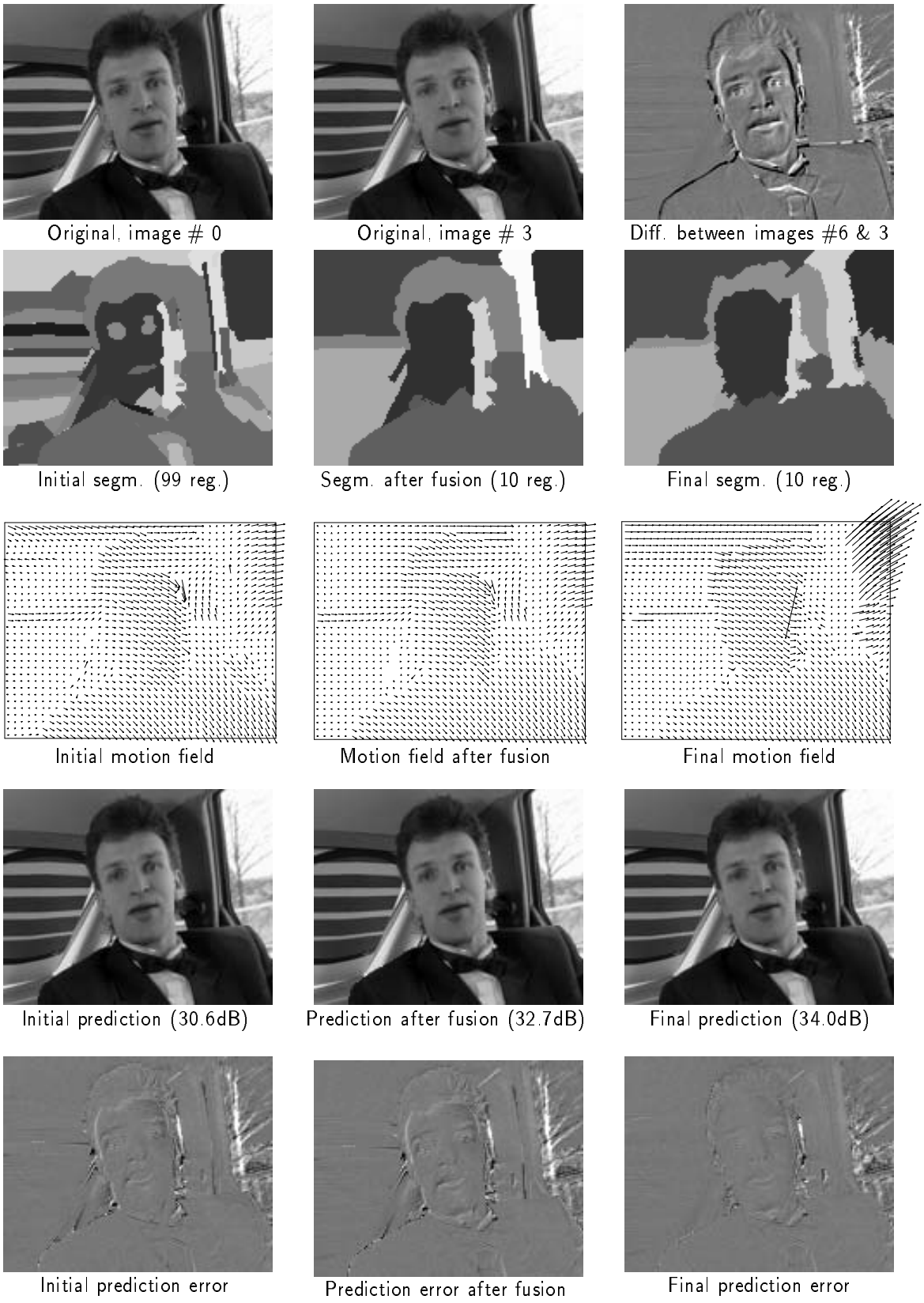


Figure 6: Two consecutive images from *"Carphone"*, their difference and the initial, after fusion and final segmentations, motion fields, prediction and error images.

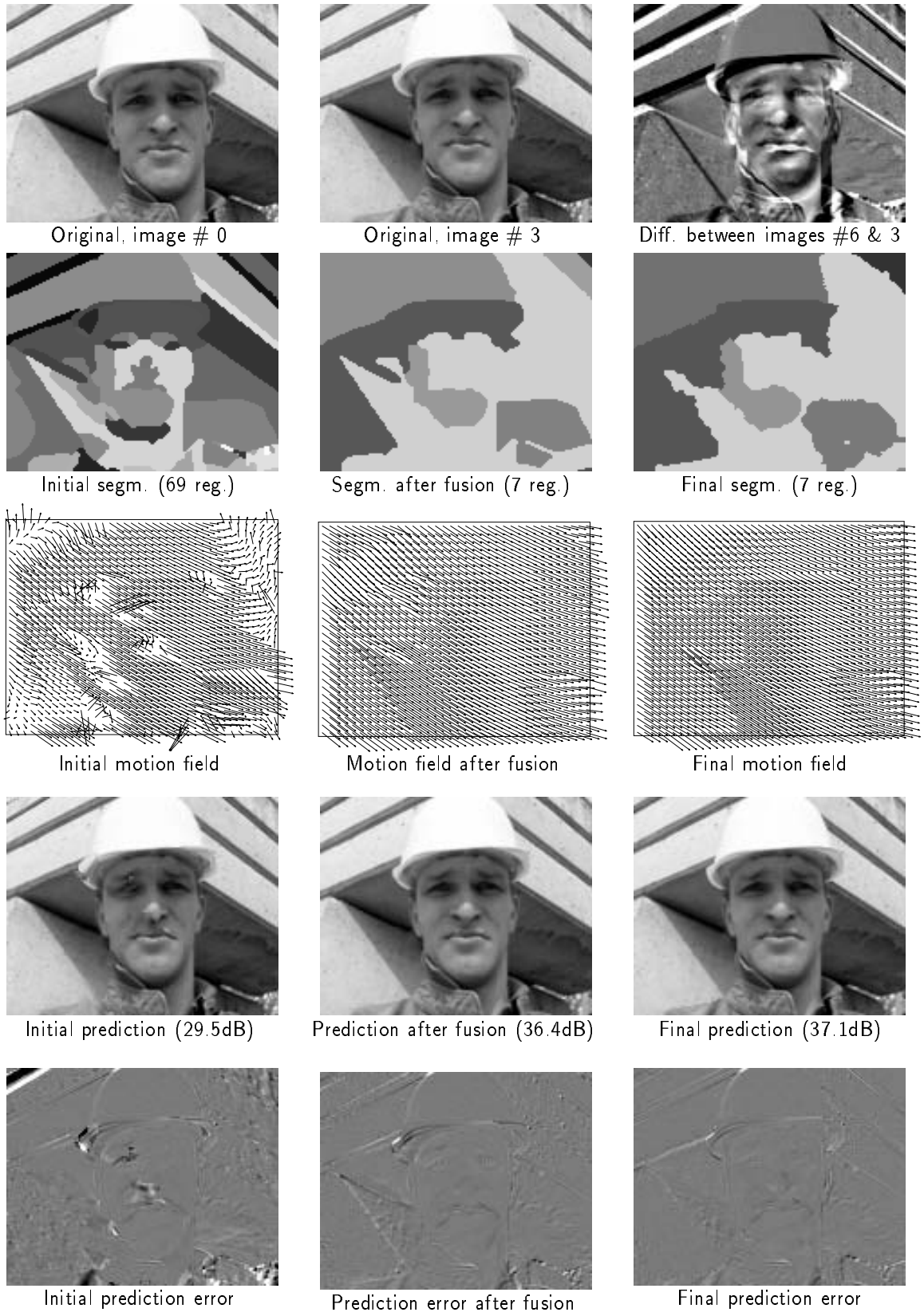


Figure 7: Two consecutive images from *"Foreman"*, their difference and the initial, after fusion and final segmentations, motion fields, prediction and error images.

In order to evaluate numerically the quality of prediction, the numbers under each predicted image give the peak prediction gain defined as:

$$PPG = 10 \log \frac{255^2}{MSPE}$$

where $MSPE$ is the mean-squared prediction error evaluated over the whole image. Clearly, subjective improvement translates into numeric gain; with every processing stage the PPG is increased for all three sequences. The overall gain ranges from 3.1dB to 7.6dB.

3.3 Efficient representation of region boundaries

Since the segmentation map $\mathcal{S}$ is an inherent part of motion representation, it needs to be encoded and transmitted. This can be done in two ways: a lossless coding, in which exact geometrical reconstruction of the regions is possible, or a lossy coding, where the reconstructed region is a close approximation to the coded one.

3.3.1 Lossless coding

We have studied lossless coding via two approaches [3]. In the first approach, we have transformed segmentation maps $\mathcal{S}$ into 4 horizontal label transitions to which we have applied entropy coding. The reversible transformation assures lossless reconstruction of $\mathcal{S}$. The transitions were Huffman-coded (based on empirical distributions), and then compressed using run-length coding and a simplified version of binary-arithmetic coding. The other approach that we have studied is chain coding applied to region boundaries. The image contours are detected using the 4-connected neighborhood (at least one neighbor different than others) and then traced point by point (clockwise or anti-clockwise). The resulting contour description consists of successive direction labels (8 labels for 8-connected neighborhoods). The assumption of smoothness of boundaries translates into long runs of the same label in the contour description. To exploit these redundancies conditional and differential coding of direction labels have been implemented. The former resulted in higher gains, and therefore only results for conditional coding are reported here.

The performance of the above algorithms for intra-frame segmentation maps is summarized in Table 1[1]. Clearly, any of the studied methods outperforms pulse-code modulation by a large margin; chain coding gives almost 100 times lower rate for this particular image. In several tests on other images, chain coding has always significantly outperformed the other two methods.

3.3.2 Lossy coding

Lossy coding of region boundaries allows only their approximate reconstruction. The method adopted is based on a polygonal approximation of region boundaries [20]. The

[1] In the sequel, bit rates are are given per frame (bits/fr), per pixel (bits/p) and per contour point (bits/cp)

"Miss America" (21 regions) "Carphone" (74 regions) "Foreman" (91 regions)

Figure 8: Intra-frame (intensity-based) segmentation maps for image #0 from the three test sequences.

	"Miss America"	*"Carphone"*	*"Foreman"*
PCM	126720	177408	177408
RLC	4053	8016	9327
BAC	2168	6871	8356
CC	1539	4604	5605

Table 1: Bit rates per frame for intra-frame segmentation maps from Fig. 8 encoded using pulse-code modulation (PCM), run-length (RLC), simplified binary-arithmetic (BAC) and chain (CC) coding. To encode all the labels from Fig. 8 using the PCM, 5 bits have been used for *"Miss America"* and 7 for the other sequences.

mean-squared pointwise error between the contour and its polygonal approximation is defined and a specified number of vertices assigned so as to minimize this error using dynamic programming. Parameter γ controls the quality of approximation; the lower the γ the closer the approximation to the original contour. Since polygonal approximation is inefficient for small regions, we use chain coding for regions of less than 30 contour points. The switching is done using a 2-bit flag, with the remaining states defining three classes that optimize polygon representation [3]. The vertex coordinates are coded differentially with respect to the preceding vertex with each class defining the maximum distance between vertices.

Table 2 shows bit rates per frame for polygonal approximation with three values of γ. Comparing this table with Table 1 we can see, as expected, that for higher γ distortion measure δ increases while the number of bits per frame decreases. Fig. 9 shows an original image, its intra-frame segmentation as well as three approximations for different values of γ. Although this is another image from the same sequence, it is very similar to image #0. As can be concluded from the table, the rates depend strongly on the accuracy of region representation by polygons, but visually little distortion can be perceived even for γ=13. This confirms our earlier observation that for optimal image quality judicious allocation of bits between intensity (color), motion and segmentation (shape) is needed.

		"Miss America"	*"Carphone"*	*"Foreman"*
γ=7	δ	0.55	0.29	0.35
	bits/fr	1223	4044	4662
γ=10	δ	1.10	0.54	0.74
	bits/fr	991	3268	3718
γ=13	δ	1.54	0.84	1.05
	bits/fr	879	2805	3254

Table 2: Bit rates per frame (bits/fr) for intra-frame segmentation maps from Fig. 8 encoded using lossy compression for three values of γ. δ is a distortion measure.

4 Results for complete video sequences

4.1 Region-based motion estimation and prediction

In Section 3.2 we have described region-based motion estimation and prediction, and we have shown some experimental results. In this section, we present results for sequences of images, we compare them with results obtained using block matching and dense motion field estimation, and finally we study the impact of lossy compression of segmentation maps on the performance of prediction.

Since the current standards such as the H.261, MPEG-1, MPEG-2 use block-based motion model, we have implemented full-search block-matching for 16×16 blocks and 1/2 pixel accuracy. Such a motion model requires transmission of only one motion vector per 256 pixels. On the other extreme of the spectrum are dense motion models that assign one vector to each pixel. Certainly, this approach substantially reduces prediction error but allocates higher bit rate to the motion information. We have implemented the dense motion field estimation according to the method described in [12] with bicubic interpolation for subpixel accuracy [11].

We have tested all three algorithms on the following parts of the test sequences:

- *"Miss America"*: frames # 3,6,9,...,57,
- *"Carphone"*: frames # 120,123,126,...,174,
- *"Foreman"*: frames # 150,153,156,...,207.

Fig. 10 shows the peak prediction gain (PPG) for the three methods. On average the region-based prediction (Section 3.2) is outperformed by the pixel-based prediction by about 2dB for *"Miss America"* and *"Carphone"*, but for *"Foreman"* the region-based prediction outperforms the pixel-based prediction by about 1dB. This may seem surprising at first, but consider the fact that motion in *"Foreman"* is very uniform due to a camera pan. To correctly compute the global translational motion, the pixel-based method would require extremely strong smoothing, but this would be detrimental in the case of multiple-object motion when motion boundaries would be blurred. On the

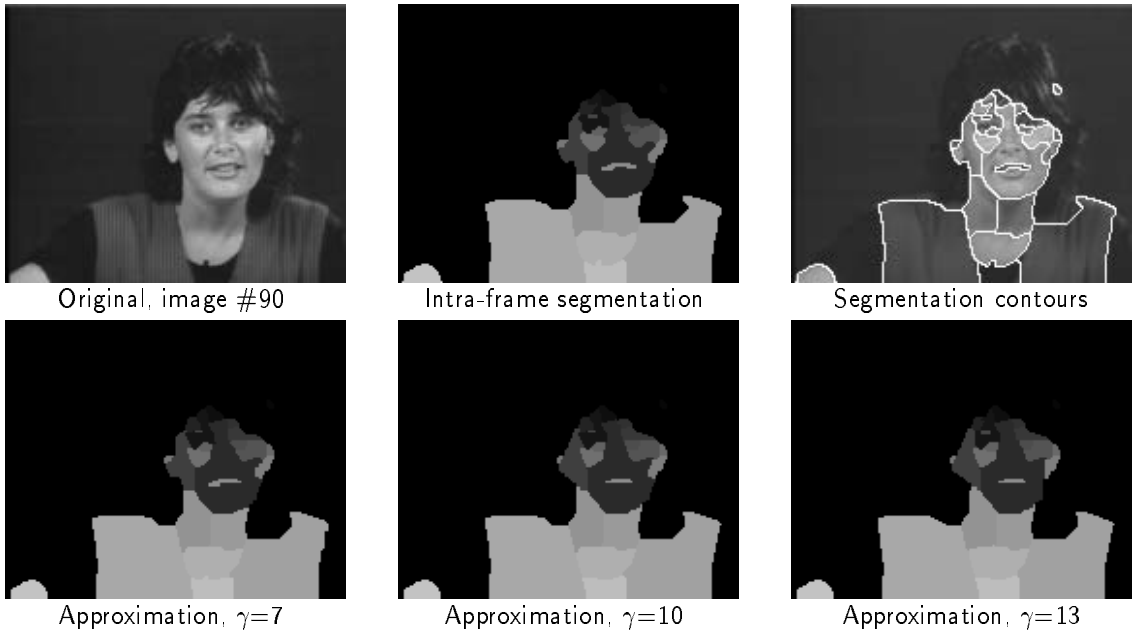


Figure 9: Original image #90 from *"Miss America"*, its intensity-based segmentation map, corresponding contours and approximations for γ=7, 10 and 13.

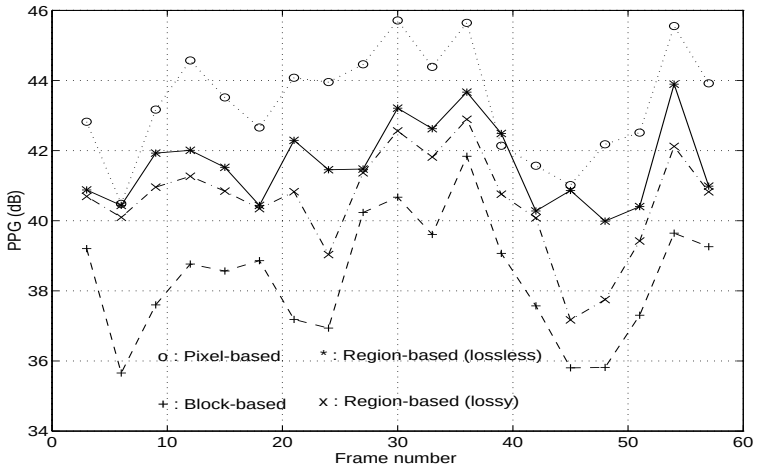
other hand, the region-based method is capable of correct computation of the global motion if sufficiently large regions are detected (which is the case here).

Note that in the pixel-based motion estimation scheme the prediction error can be made very small by relaxing the motion smoothness constraint [12]. To make the comparison fair, prediction errors should be evaluated for motion information requiring the same rate in all three cases. Since we have no suitable coding scheme for motion information, we chose a high degree of smoothness in the pixel-based scheme that results in a relatively large error.

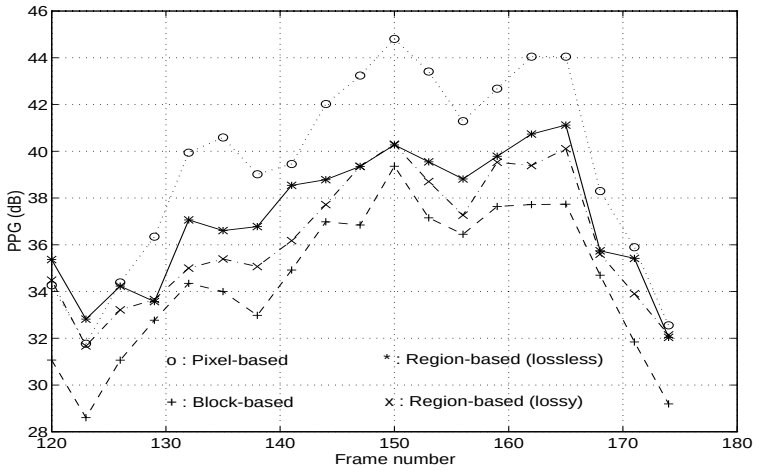
At the same time the region-based prediction outperforms the block-based prediction by about 3-4 dB for all three sequences.

In Figure 10 the region-based PPG marked as "lossless" has been obtained under the assumption that segmentation is transmitted with no loss of information. We have also applied lossy compression (Section 3.3) to the segmentation maps $\mathcal{S}$ and subsequently we have recomputed the motion parameters Φ. As can be seen from Fig. 10, this approximation ("lossy") causes an average PPG drop of about 1dB for the three sequences when compared with lossy compression.

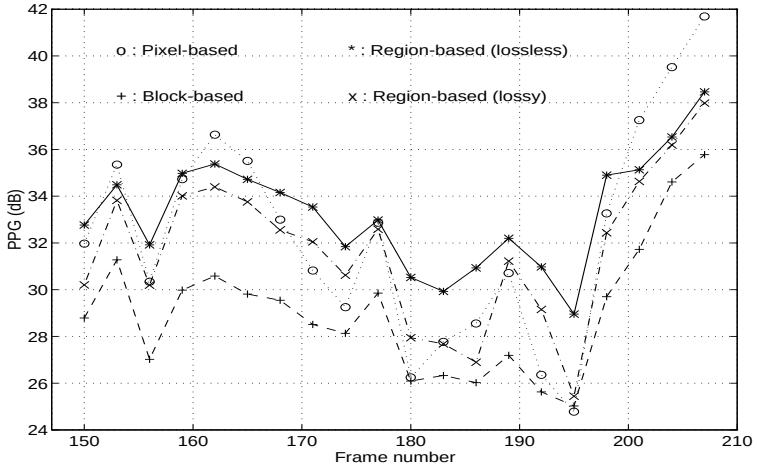
Fig. 11 shows some results for image #171 of the sequence *"Carphone"*. Note the significant simplification of the intensity-based partition by fusion and boundary adjustment. Although the final partition is far from perfect, it corresponds reasonably well to the motion of different objects in the scene. The most obvious difference between motion fields is in the car window, but since almost no intensity gradient is present there it has little impact on the prediction. The region-based prediction errors are concentrated mostly in the vicinity of region boundaries. These errors are



"Miss America"



"Carphone"



"Foreman"

Figure 10: Peak prediction gain of the motion-compensated prediction error for block-based, pixel-based and region-based (lossless and lossy) motion estimates for the three test sequences.

less visible in the prediction images due to the spatial masking of the HVS, and it is likely that they will go unnoticed in the encoded images. The block-based model results in the largest prediction error, especially in the area of the mouth and chin (block visibility). These distortions will be most likely visible in the encoded images if a very limited bit budget is available. This is certainly the case for the H.261 standard and for the TMN2 model at rates below 20 kb/s.

4.2 Representation of region-boundaries

In Section 3.3 we have presented compression results for a single frame of intra-frame segmentation maps. Here we present results for a sequence of frames of each test image.

First, we have tested the intra-frame (intensity-based) segmentation on the following parts of the test sequences:

- *"Miss America"*: frames # 0,5,10,...,145,
- *"Carphone"*: frames # 0,1,2,...,19,120,123,...,177,
- *"Foreman"*: frames # 0,1,2,...,19,150,153,...,207.

The subsampling by 5 in *"Miss America"* and by 3 in the second part of other sequences was done to simulate typical processing applied to videoconferencing material before encoding. Table 3 gives the average number of regions and contour points in each of the sequences, while Tables 4 and 5 show average rates for lossless and lossy compression, respectively.

	"Miss America"	*"Carphone"*	*"Foreman"*
Regions	20	55	75
Contour points	830	2425	2760

Table 3: Average number of regions and contour points for intra-frame segmentations of the test sequences.

In terms of the number of bits per contour point (bits/cp), the run-length coding and the simplified binary arithmetic coding perform similarly; the increase in the case of run-length coding for *"Miss America"* is due to the substantially longer runs of "no transition" states. Clearly, chain coding outperforms the other two lossless coding techniques in terms of the number of bits/fr and bits/cp, however, the latter number excludes the coding of starting points. Note that the number of bits per contour point is within 20% of the estimated entropy rate for chain codes (under the assumption of stationarity and first-order Markov statistics). As for lossy coding (Table 5), the rates can be reduced below those of chain coding even for high-quality approximation (γ=7). As mentioned before since the rate allocated to segmentation maps changes substantially with the change of γ while the HVS-perceived quality of approximation

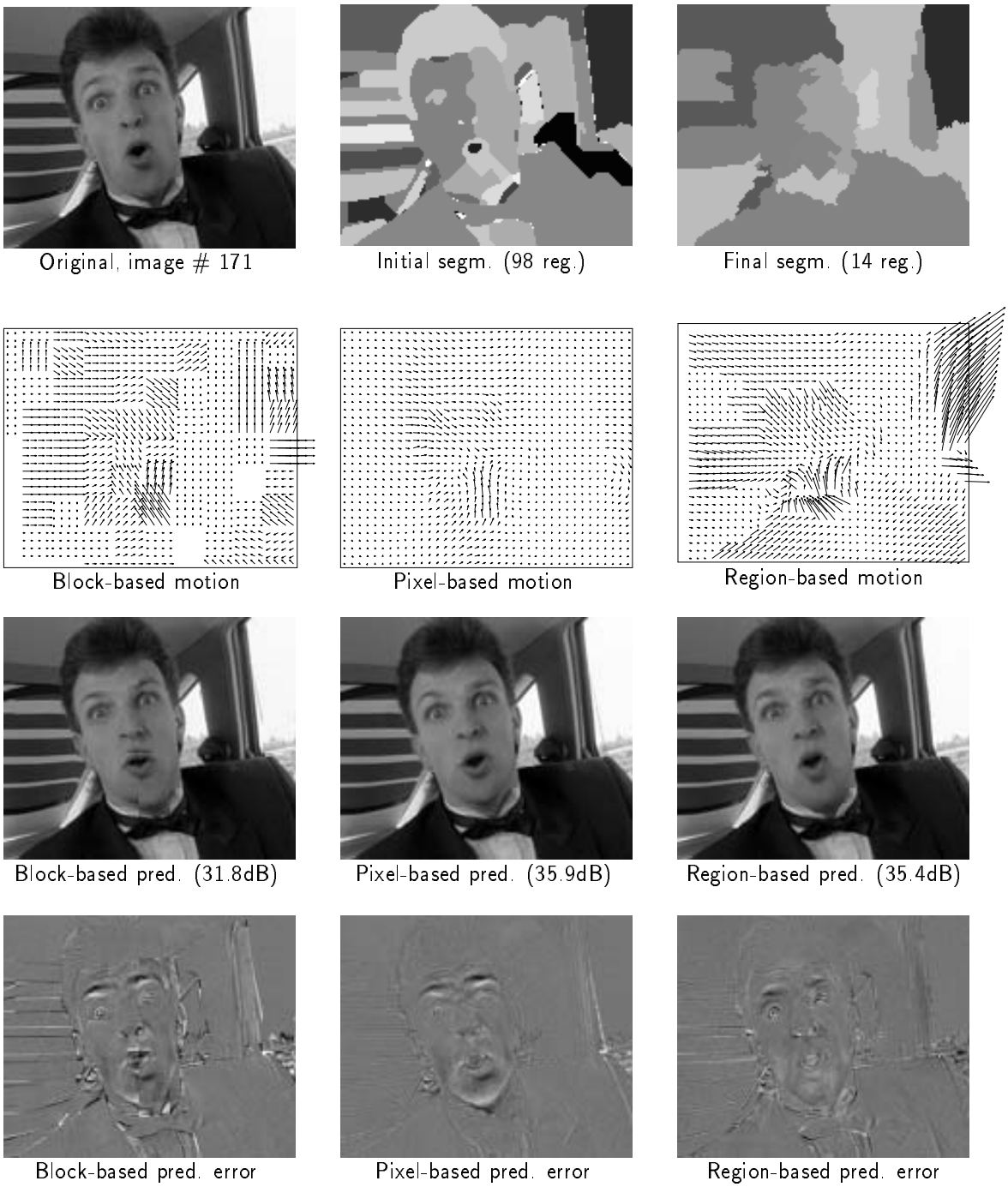


Figure 11: Original image # 171 from *"Carphone"*, initial and final segmentation maps, motion fields, motion-compensated predicted images and prediction errors.

		"Miss America"	*"Carphone"*	*"Foreman"*
RLC	bits/fr	4000	7230	9325
	bits/p	0.158	0.285	0.368
	bits/cp	4.80	2.98	3.38
BAC	bits/fr	2135	5885	8150
	bits/p	0.084	0.232	0.322
	bits/cp	2.57	2.43	2.95
CC	bits/fr	1447	4010	5090
	bits/p	0.057	0.158	0.201
	bits/cp*	1.36	1.29	1.41
	entropy	1.20	1.16	1.24

Table 4: Average rates over sequences of intra-frame segmentations encoded using the three lossless methods. *Starting points of chains are not accounted for. The entropy is an estimate of the entropy rate under the assumption that the process consisting of direction chain codes is stationary and first-order Markov.

		"Miss America"	*"Carphone"*	*"Foreman"*
γ=7	δ	0.47	0.31	0.36
	bits/fr	1234	3485	4257
	bits/p	0.049	0.138	0.168
	bits/cp	1.48	1.44	1.54
γ=10	δ	0.95	0.61	0.72
	bits/fr	1012	2805	3392
	bits/p	0.040	0.111	0.133
	bits/cp	1.22	1.16	1.23
γ=13	δ	1.45	0.91	1.10
	bits/fr	900	2440	2940
	bits/p	0.036	0.096	0.116
	bits/cp	1.08	1.01	1.06

Table 5: Average rates over sequences of intra-frame segmentations encoded using the lossy method for three values of γ.

changes very little, the only way to assure overall maximum image quality is to select optimal bit allocation between texture, motion and shape.

As shown in Fig. 1 coding of region boundaries is also needed in the inter-frame (motion-based) coding branch of the proposed system. Since usually image partition based on motion is quite different statistically than the one based on intensity only (more regions in intensity-based segmentation; not everything moves), we have tested lossless and lossy compression on motion-based (inter-frame) segmentation maps. We have used the following sequences in our tests:

- *"Miss America"*: frames # 3,6,9,...,57,
- *"Carphone"*: frames # 120,123,...,177,
- *"Foreman"*: frames # 150,153,...,207.

The average numbers of regions and contour points are given in Table 6, while average bit rates for the lossless and lossy compression are shown in Tables 7 and 8, respectively. Clearly, motion-based partitions are less complex than their intensity-based counterparts; there are fewer regions and fewer contour points. Consequently, the rates for motion-based segmentation are lower. In terms of the number of bits/cp, lossless techniques perform similarly for both types of segmentation; the increase in the case of run-length coding for *"Carphone"* and *"Foreman"* is due to the substantially longer runs of "no transition" states. Again, chain coding outperforms the other lossless techniques. A very interesting alternative, however, is lossy coding since it is capable of approaching rates of less than 1kb/fr (10kb/s at 10Hz). These rates, however, depend strongly on the accuracy of region representation by polygons. Fig. 12 shows an example of final motion-based segmentations and their approximations by polygons (γ=10).

	"Miss America"	*"Carphone"*	*"Foreman"*
Regions	5	10	15
Contour points	540	875	1115

Table 6: Average number of regions and contour points for motion-based segmentations of the test sequences.

5 Demonstration of results

In order to present visually some of the results obtained in the scope of this work, we have prepared three demonstrations of short video sequences. These demonstrations can be viewed in the Visual Communications group laboratory on the *Viewstore 6000* image sequencer. The following video demonstrations have been prepared for each of the test sequences, i.e., *"Miss America"*,*"Carphone"* and *"Foreman"*:

		"Miss America"	*"Carphone"*	*"Foreman"*
RLC	bits/fr	2622	2851	4685
	bits/p	0.103	0.112	0.185
	bits/cp	4.85	3.26	4.20
BAC	bits/fr	1380	2200	3081
	bits/p	0.054	0.087	0.122
	bits/cp	2.55	2.51	2.76
CC	bits/fr	767	1380	1810
	bits/p	0.030	0.054	0.071
	bits/cp*	1.27	1.39	1.40
	entropy	1.07	1.16	1.17

Table 7: Average rates over sequences of motion-based segmentations encoded using the three lossless methods. *Starting points of chains are not accounted for. The entropy is an estimate of the entropy rate under the assumption that the process is stationary and first-order Markov.

		"Miss America"	*"Carphone"*	*"Foreman"*
γ=7	δ	0.69	1.75	1.28
	bits/fr	482	831	1285
	bits/p	0.019	0.033	0.051
	bits/cp	0.89	0.95	1.15
γ=10	δ	1.05	2.89	2.10
	bits/fr	383	725	1140
	bits/p	0.015	0.029	0.045
	bits/cp	0.71	0.83	1.02
γ=13	δ	1.43	3.92	2.88
	bits/fr	331	635	995
	bits/p	0.013	0.025	0.039
	bits/cp	0.62	0.72	0.89

Table 8: Average rates over sequences of motion-based segmentations encoded using the lossy method for three values of γ.

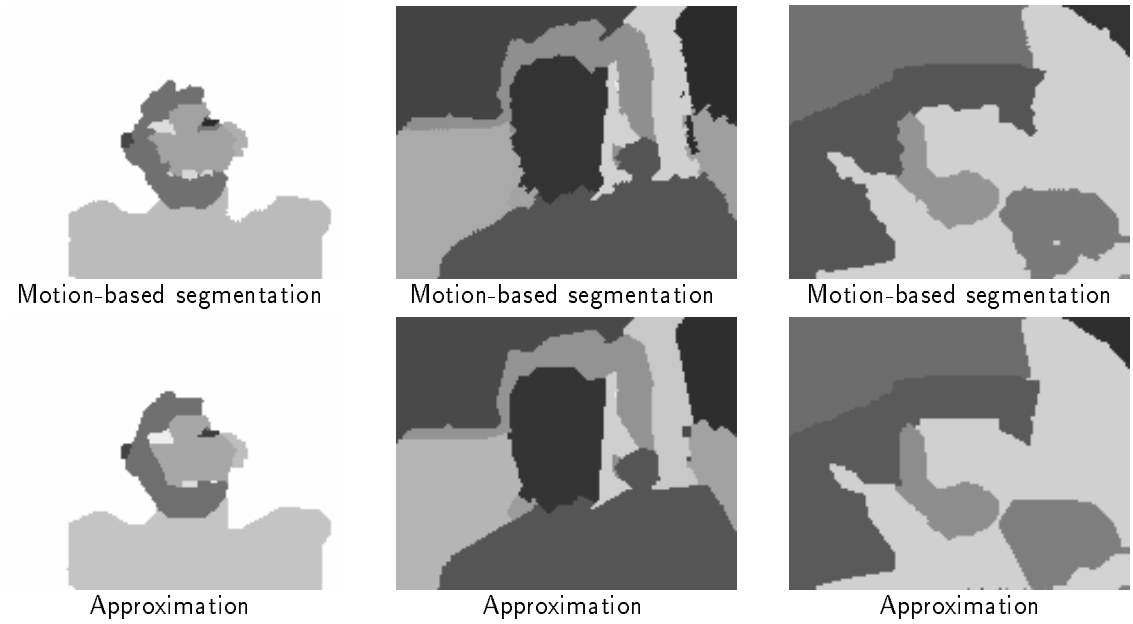


Figure 12: Motion-based (final) segmentation maps from Figs. 5, 6 and 7 after lossy compression. Average bit-rates are given in Table 8.

- *Intra-frame (intensity-based) segmentation via the MDL algorithm* (Section 3.1). Sequences shown:
  - original sequence,
  - MDL intra-frame segmentation sequence.
- *Region-based segmentation and estimation of motion* (Section 3.2). Two groups of sequences shown:
  - original sequence,
  - motion-compensated prediction sequence for block-based motion model (block matching),
  - motion-compensated prediction sequence for pixel-based motion model,
  - motion-compensated prediction sequence for region-based motion model,

  and
  - inter-frame (motion-based) segmentation sequence,
  - motion-compensated prediction error sequence for block-based motion model (block matching),
  - motion-compensated prediction error sequence for pixel-based motion model,
  - motion-compensated prediction error sequence for region-based motion model.

No video demonstration is provided for the results from Sections 2 and 3.3. In the first case, sequences have been encoded in order to evaluate their $PSNR$ and to compare the rate-distortion performance of two motion estimation algorithms. In the second case, for lossless compression no data loss occurs while for the high-quality

(γ=7 or 10) lossy compression, that we are interested in, the differences between the original and approximate segmentations are so small that it would be difficult to see them on the screen.

6 Discussion

As anticipated in the proposal we have continued to study standard block-based compression schemes, while at the same time we have looked at more general schemes that allow natural partitioning of video sequences into regions.

As the results from Section 2 indicate, gains attained by rate-based motion estimation within the H.261 coding syntax are about 1.0dB and this only at higher rates (higher quality). Thus, the potential for improvement within the H.261 standard is very limited, especially at lower rates. Since new video applications, such as rapidly growing desktop videoconferencing and awaited mobile communications, clearly aim at such rates, a modification of the H.261 standard seems to be unavoidable. To date several attempts at relatively small modification of the standard have been made. For example, motion compensation based on overlapped blocks [19] and on vector field interpolation [18] have been proposed to reduce the "blockiness" of image intensities. Although both methods show a substantial improvement in the quality of prediction images, especially subjectively, it is not clear at this point whether the achieved gains will be sufficient for application at very low rates, e.g., of less than 16 kb/s. While some of the above concepts have found way into the new H.263 proposal, further work is needed to gain insight into the performance limits of various H.261-like schemes. For example, recent work on block splitting to achieve more precise motion estimates [7] goes in that direction.

Despite the initial activity within the MPEG-4 group to propose object- or region-based coding as the future very low bit rate video coding standard, it remains to be demonstrated that such schemes form a viable alternative to H.261-like algorithms. Above, we have proposed an approach to reliable intensity-based (intra-frame) segmentation and motion estimation/segmentation, two important elements of the scheme from Figure 1. As the results demonstrate, the region-based prediction outperforms the block-based prediction by 3-4dB, while it is worse than the pixel-based prediction by about 2dB for fairly localized motion. For global motion, however, the region-based method is better by about 1dB. Although these results cannot be easily extrapolated to arbitrary data, we expect that for large areas of consistent motion the region-based prediction should perform very well. We believe that this approach is very promising, but we acknowledge the need for its further refinement to account for temporal consistency of regions [8].

We have also evaluated several algorithms for the compression of intensity- and motion-based segmentation maps. Although the best compression can be achieved with lossy coding, in order to maintain a good correspondence (small error) between the original and approximated regions, the rate of about 0.5-1kb/frame is still needed for QCIF sequences; this may be too high a rate for very low bit rate applications

(transmission at 10Hz would require 5-10kb/s just for the shape information). Since our scheme did not account for temporal correlation of the segmentation maps, this rate may be reduced by eliminating the temporal redundancy [21]. This is related to the temporal consistency of regions mentioned above and studied in [8].

Of the remaining blocks from diagram in Fig. 1, the coding of motion parameters has not been tackled yet, while we have done some preliminary work (outside of this project) on the intra-frame coding of arbitrarily-shaped regions [22]. The inter-frame coding remains to be studied and again it is closely related to the temporal consistency of regions.

7 Further directions

In continuation of the work carried out under phase II of the project we can see the following directions:

- traditional block-based coding:
  - evaluate the new very low bit rate video coding proposal H.263,
  - study improved motion models (e.g., segmentation) for traditional schemes,
- region-based coding:
  - study ways of exploiting temporal consistency of regions in a video sequence to reduce bit rate allocated to the transmission of segmentation maps (shape),
  - study inter-frame coding, e.g., motion-compensated two-frame versus multiple-frame prediction,
  - study the coding of motion parameters.

References

[1] CCITT, *Recommendation H.261: Video codec for audiovisual services at p x 64 kbits/s*, 1990.

[2] S.-F. Chang and D. Messerschmitt, "Transform coding of arbitrarily-shaped images segments," in *Proc. ACM Multimedia Conf.*, pp. 83–90, Aug. 1993.

[3] J.-B. Chartier, "Représentation compacte des cartes de segmentation dans les séquences d'images: étude comparative," Master's thesis, INRS-Télécommunications, Mar. 1995.

[4] H. Chen, M. Civanlar, and B. Haskell, "A block transform coder for arbitrarily-shaped image segments," in *Proc. IEEE Int. Conf. Image Processing*, pp. 85–89, Nov. 1994.

[5] A. P. Dempster, N. M. Laird, and D. P. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Roy. Stat. Soc.*, vol. B 39, no. 1, pp. 1–38, 1977.

[6] N. Diehl, "Object-oriented motion estimation and segmentation in image sequences," *Signal Process., Image Commun.*, vol. 3, pp. 23–56, Feb. 1991.

[7] F. Dufaux, I. Moccagatta, F. Moscheni, and H. Nicolas, "Vector quantization-based motion field segmentation under the entropy criterion," *J. Vis. Commun. Image Represent.*, vol. 5, pp. 356–369, Dec. 1994.

[8] V. Garcia-Garduño, C. Labit, and L. Bonnaud, "Temporal linking of motion-based segmentation for object-oriented image sequence coding," in *Signal Process. VII: Theories and Applications (Proc. Seventh European Signal Process. Conf.)*, pp. 147–150, Sept. 1994.

[9] M. Gilge, T. Engelhardt, and R. Mehlan, "Coding of arbitrarily shaped image segments based on a generalized orthogonal transform," *Signal Process., Image Commun.*, vol. 1, pp. 153–180, Oct. 1989.

[10] J. Hutchinson, C. Koch, J. Luo, and C. Mead, "Computing motion using analog and binary resistive networks," *Computer*, vol. 21, pp. 52–63, Mar. 1988.

[11] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-29, pp. 1153–1160, Dec. 1981.

[12] J. Konrad and E. Dubois, "Comparison of stochastic and deterministic solution methods in Bayesian estimation of 2D motion," *Image Vis. Comput.*, vol. 9, pp. 215–228, Aug. 1991.

[13] J. Konrad and E. Dubois, "Bayesian estimation of motion vector fields," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-14, pp. 910–927, Sept. 1992.

[14] J. Konrad, A.-R. Mansouri, E. Dubois, V.-N. Dang, and J.-B. Chartier, "On motion modeling and estimation for very low bit rate video coding," in *Proc. SPIE Visual Communications and Image Process.*, vol. 2501, pp. 262–273, May 1995.

[15] Y. Leclerc, "Constructing simple stable descriptions for image partitioning," *Intern. J. Comput. Vis.*, vol. 3, pp. 73–102, 1989.

[16] A.-R. Mansouri and J. Konrad, "A comparative evaluation of the H.261 and the SIM-3 video coders," Tech. Rep. 94–30, INRS-Télécommunications, Oct. 1994.

[17] H. G. Musmann, M. Hötter, and J. Ostermann, "Object-oriented analysis-synthesis coding of moving images," *Signal Process., Image Commun.*, vol. 1, pp. 117–138, Oct. 1989.

[18] J. Niewęgłowski and P. Haavisto, "Motion vector field reconstruction for predictive video sequence coding," in *Int. Workshop on Coding Techniques for Very low Bit-rate Video, VLBV'94*, Apr. 1994.

[19] M. Orchard and G. Sullivan, "Overlapped block motion compensation: an estimation-theoretic approach," *IEEE Trans. Image Process.*, vol. 3, pp. 693–699, Sept. 1994.

[20] J. Perez and E. Vidal, "Optimum polygonal approximation of digitized curves," *Pattern Recognit. Lett.*, vol. 15, pp. 743–750, Aug. 1994.

[21] C. Stiller, "Object oriented video coding employing dense motion fields," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, pp. V.273–V.276, Apr. 1994.

[22] C. Stiller and J. Konrad, "Eigentransforms for region-based image processing," in *Proc. Int. Conf. on Consumer Electronics*, pp. 286–287, June 1995.

[23] C. Stiller and J. Konrad, "A region-adaptive transform based on a stochastic model," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 1995 (in print).