Low Bit-Rate Coding of Image Sequences Using Adaptive Regions of Interest

Nikolaos Doulamis, Anastasios Doulamis, Dimitrios Kalogeras, and Stefanos Kollias

***Abstract*— An adaptive algorithm for extracting foreground objects from background in videophone or videoconference applications is presented in this paper. The algorithm uses a neural network architecture that classifies the video frames in regions-of-interest (ROI) and non-ROI areas, also being able to automatically adapt its performance to scene changes. The algorithm is incorporated in motion-compensated discrete cosine transform (MC–DCT)-based coding schemes, allocating more bits to ROI than to non-ROI areas. Simulation results are presented, using the Claire and Trevor sequences, which show reconstructed images of better quality, as well as signal-to-noise ratio improvements of about 1.4 dB, compared to those achieved by standard MC–DCT encoders.**

***Index Terms*— Low bit-rate coding, MC-DCT-based coding schemes, neural networks, regions of interest.**

I. Introduction

In previous years, efforts for image sequence coding at different bit rates have stimulated the generation of various standards, such as MPEG-1 and MPEG-2 [1]. Transmission of video signals through conventional or mobile telephony, however, requires, on the one hand, high compression ratios, and on the other hand, preservation of good picture quality. In this framework, the H.263 standard has been generated for improving the video quality provided by the former standards at bit rates lower than 64 kbit/s [2]. Moreover, MPEG-4 [3], [4] aims at developing algorithms for audio–visual coding in multimedia applications, which allow high compression ratios, interactivity, universal accessibility, and portability of audio and video content. It adopts the concept of video objects (VO's) and video object planes (VOP's) of arbitrary shape, permitting separate decoding and composition of them. Consequently, some video objects are decoded and presented to the viewers, while some others may be substituted synthetic ones.

Excluding video games or graphics applications, where object segmentation is *a priori* available, extraction of video objects is a rather hard task. Segmentation techniques based on spatial and/or motion homogeneity criteria have been proposed for this purpose [5], [6]. Nevertheless, a physical object, such as a person in a scene, contains regions with different color and texture (e.g., head, hair, clothes' color) which can belong to different segments according to such homogeneity criteria. Moreover, physical objects are not only the moving objects in a scene; only a part of a physical object may be moving during a frame period as, for example, when a speaker moves his/her head and hands, while keeping the rest of his/her body still [7].

In this paper, we propose a new technique for extracting foreground VOP's, e.g., head and shoulder parts of speakers, from background ones in videophone or videoconference applications. Extraction of foreground VOP's, which are called regions-of-interest (ROI) in the following, is based on a two-level neural network classifier that is described in Section II. The first level of the classifier provides an approximate classification of VOP's in foreground and background ones, while the second level improves the obtained classification accuracy and adapts to scene changes, based on well-known object connectivity criteria and on an automatic retraining procedure. The proposed method is computationally efficient, especially when compared to conventional segmentation techniques.

In Section III, the proposed technique is combined with the MPEG-1 video coding algorithm; other coding standards, such as H.263, could similarly take advantage of it. The rate control of the MPEG-1 algorithm is modified so that the quality of reconstructed foreground VOP's is higher than that of background ones. This is achieved by applying coarser quantization to the latter parts of the video frames and finer to the former ones. Simulation results are presented in Section IV, while conclusions and further work are given in Section V of the paper.

II. Adaptive ROI Selection Using Neural Networks

A neural-network-based scheme is applied to image sequences for extracting foreground (ROI) VOP's from background ones. Each frame is first divided into rectangular blocks of, say, 8×8 pixels. Appropriate features are then extracted from each block and used as inputs to a neural classifier, which determines the class (foreground/background) the respective block belongs to. A binary segmentation mask is formed next, including the classifier binary outputs over all blocks of the frame, which specifies the locations of foreground and background VOP's in the frame, at block resolution. This information can be included in MPEG-1 or H.263 compatible encoders, in order to improve their performance at low bit rates, by allocating more bits to foreground VOP's than to background, as described in the following section. In the case of "object-layered" encoders, like MPEG-4, segmentation of video objects at pixel resolution can be achieved through postprocessing of the segmentation mask. In particular, blocks which include object boundaries are first selected as those blocks for which at least one neighboring block does not belong in the same VOP category. Segmentation at pixel level is then achieved through edge detection within these "boundary" blocks, while preserving the continuity of VOP's between adjacent blocks.

Manuscript received September 28, 1995; revised February 20, 1996 and April 20, 1998. This paper was recommended by Associate Editor T. Sikora.

The authors are with the Department of Electrical and Computer Engineering, National Technical University of Athens, 157 73 Zografou, Athens, Greece.

Publisher Item Identifier S 1051-8215(98)09284-2.

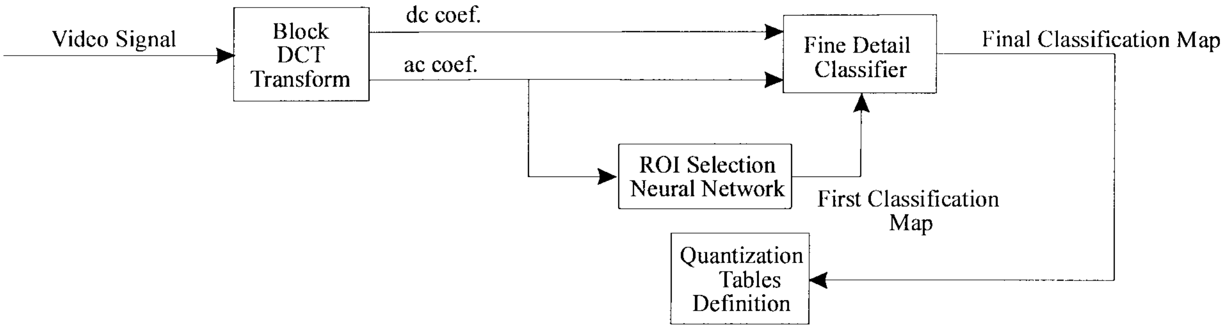


Fig. 1. Two-level neural network architecture.

The above-described neural network classifier has a two-level architecture as shown in Fig. 1. The first level consists of a feedforward neural network [8], which generates the segmentation mask using the features extracted from each image block. The second level, also using a feedforward network, can further improve the obtained segmentation, exploiting object continuity in the segmentation mask provided by the first network and possibly using additional features.

A zig-zag scanned portion of the ac coefficients of the discrete cosine transform (DCT) of each block comprises the features used by the first network. The number of input nodes of the network equals the number of the ac coefficients used; in the case of binary (foreground/background) classification, the network has two outputs, corresponding to the two possible classes. A set of characteristic examples of blocks belonging to foreground (e.g., eyes, mouth, hair, clothes) and to background is selected and used to provide the features for training the network. Since these features are the ac DCT coefficients, the network learns to perform classification of blocks in the frequency domain; it then operates as shown in Fig. 1. It is, however, possible even for a well-trained network not to perform satisfactorily when the operational environment is different from the initial training conditions, e.g., due to a change of luminosity or color of clothes or position of persons. In such cases, the network can misclassify some blocks, thus providing an approximate, not accurate segmentation of the frame in foreground and background VOP's.

For this reason, a mechanism is introduced which detects, in each frame, whether or not such a change of environment takes place. A retraining procedure is used in the former case, which overcomes the above-mentioned misclassifications. Let us first assume that a frame, in which such a change exists, has been detected. We wish to select those blocks of the frame which have been "correctly" classified to a foreground or to a background VOP by the network. The criterion we use is local connectivity of VOP's, i.e., the fact that all blocks within a VOP should belong to the same category. According to this criterion we select those blocks, all neighbors of which, in a window of 3 × 3 blocks around them, belong to the same class, according to the segmentation mask provided by the network. The selected blocks form a new training data set which is used next for training a second feedforward neural network to perform the classification task. This latter network then will be applied to the same frame, from which the new training data have been selected. Additional features which are characteristics of the specific frame may consequently be used for training the network, providing it with the ability to classify image blocks which might have been erroneously classified by the first network. The most appropriate additional feature is color, provided by the dc DCT coefficient of the three color components of each block; this is because, within the same scene, color information changes slowly with time. Consequently, after being trained with blocks from the first frame of the new scene, the network will be able to generalize its good performance in the following frames of the scene until a new change of the environment is detected.

Let us now present the decision mechanism. During operation of the proposed two-level scheme, the corresponding segmentation masks provided by the first and second networks will be slightly different; their main difference will be in some misclassified foreground and/or background blocks. When a new change of the environment occurs, the second network, having being trained with color features of the previous scene, will fail. The first network will still, however, provide an approximate segmentation mask since it has been trained with the frequency content, and not with the specific color conditions. Consequently, in this case, the difference between the segmentation masks provided by the two networks will be large. Automatic detection of such changes is, therefore, possible through a continuous comparison of the masks provided by the two networks at each frame of the sequence. Retraining of the second network will automatically be performed, using the selected (as described above) data whenever such a change is detected.

The learning vector quantization (LVQ) algorithm [8] has been chosen and used for training both networks since it can be implemented in real time, while giving accurate results. LVQ considers the network weights as representatives of the desired classes. In the nth iteration, the algorithm compares the corresponding input, say $\boldsymbol{t}$, with the network weights to find the weight, say $\boldsymbol{w}_c$, which is closer to $\boldsymbol{t}$. If the classes of $\boldsymbol{t}$ and $\boldsymbol{w}_c$ are the same, then $\boldsymbol{w}_c$ is moved closer to $\boldsymbol{t}$; otherwise, it is moved far from it. In particular, the weight $\boldsymbol{w}_c$ is adjusted as follows.

- If classes of $\boldsymbol{t}$ and $\boldsymbol{w}_c$ agree, then

$$\boldsymbol{w}_c(n+1) = \boldsymbol{w}_c(n) + a(n)\,(\boldsymbol{t} - \boldsymbol{w}_c(n)). \qquad (1)$$

- Otherwise,

$$\boldsymbol{w}_c(n+1) = \boldsymbol{w}_c(n) - a(n)\,(\boldsymbol{t} - \boldsymbol{w}_c(n)) \qquad (2)$$

TABLE I
AVERAGE PSNR PRODUCED BY THE ROI MC–DCT ALGORITHM COMPARED TO THAT PROVIDED BY CONVENTIONAL MPEG-1 AND PERCENTAGE OF BITS ALLOCATED TO ROI BLOCKS IN THE CASES OF CLAIRE/TREVOR SEQUENCES

Claire Sequence

	ROI MC-DCT Encoder			MC-DCT Encoder			ROI / Total Bits	
Bitrate (Kbits/s)	Average PSNR	Intra Frame PSNR	Inter Frame PSNR	Average PSNR	Intra Frame PSNR	Inter Frame PSNR	ROI/Total lBits (Intra Frames)	ROI/Total Bits (Inter Frames)
16	34.4	34.6	34.5	33.0	32.8	33.0	49.0%	16.1%
20	34.7	34.8	34.7	33.4	33.0	33.4	49.7%	20.0%
25	35.3	35.1	35.3	34.0	33.5	34.1	50.3%	26.4%
30	36.1	36.1	36.1	34.8	34.2	34.9	51.1%	31.8%
35	37.1	37.5	37.0	35.5	34.8	35.6	52.1%	32.5%
40	38.0	38.3	38.0	36.1	35.3	36.2	53.4%	33.9%
45	38.5	38.5	38.5	36.8	35.9	36.9	54.2%	34.8%
50	39.0	38.6	39.0	37.3	36.4	37.4	54.9%	38.6%
55	39.3	38.7	39.4	37.7	36.8	37.8	55.7%	42.9%
60	39.7	38.9	39.8	38.0	37.0	38.1	56.2%	43.9%

Trevor Sequence

	ROI MC-DCT Encoder			MC-DCT Encoder			ROI / Total Bits	
Bitrate (Kbits/s)	Average PSNR	Intra Frame PSNR	Inter frame PSNR	Average PSNR	Intra frame PSNR	Inter Frame PSNR	ROI/Total Bits (Intra Frames)	ROI/Total Bits (Inter Frames)
16	29.5	29.5	29.5	28.2	28.8	28.1	62.4%	22.1%
20	29.7	29.6	29.7	28.3	28.9	28.2	63.0%	25.3%
25	29.9	29.8	29.9	28.7	29.2	28.6	63.4%	27.1%
30	30.2	30.2	30.2	29.0	29.3	29.0	64.1%	32.9%
35	30.6	30.5	30.6	29.2	29.5	29.2	65.0%	34.4%
40	31.1	30.8	31.0	29.7	29.7	29.7	66,3%	38.4%
45	31.3	31.3	31.3	30.1	29.9	30.1	67.4%	42.1%
50	31.6	31.5	31.6	30.4	30.2	30.4	68,0%	44.5%
55	31.8	31.7	31.8	30.7	30.3	30.7	68.6%	46.4%
60	32.2	31.9	32.2	30.9	30.6	30.9	69.1%	47.4%

while the other weights are not modified and $a(n)$ is a learning parameter with $0 < a(n) < 1$. It is generally desirable that the learning parameter $a(n)$ decreases monotonically with the number of iterations n. After several passes through the input data, the network weights converge and the training is completed.

III. THE ROI-BASED MC–DCT CODER

Direct application of the MPEG-1 algorithm, for example, to low bit-rate video coding would imply bit allocation strategies which impose coarse quantization to the whole image. This, however, would heavily deteriorate the quality of the video frames both in the foreground and background VOP's.

In the proposed approach, we have modified the MPEG rate control, preserving its compatibility to the MPEG-1 algorithm so as to allocate more bits to foreground objects, where the human visual system is more sensitive, than to background ones. In standard MPEG-1 coding and for given target bit rates, frame rates, and image sizes, rate control estimates the number of bits to be allocated to coding of I, P, and B frames within each group of pictures (GOP). In the proposed ROI-based MC–DCT encoder, the rate control mechanism also exploits information provided by the ROI selection module, which indicates whether each block belongs to a foreground or a background VOP. Based on this information, it computes the number of bits that should be allocated to the foreground and background VOP's of each frame, producing a higher bit rate for foreground VOP's than conventional MPEG-1, by reallocating bits from background to foreground.

In very low bit-rate cases, however, when it may not be possible to reallocate bits, the ROI-based MC–DCT algorithm still forces the foreground areas to be coded at a higher rate than the estimated one in order to preserve the picture quality. This causes an increase of the total bit rate, which starts being evident from the beginning of of each GOP, i.e., when coding intraframes. In the next frame within each GOP (interframe coding), the rate control mechanism perceives the increase of the total bit rate and uses a higher quantization factor (allocating fewer bits) to motion-estimated prediction error so as to produce the required total bit rate. Decoded foreground VOP's are still of better quality than the ones produced by conventional MPEG-1 because the motion-compensated prediction errors in foreground VOP's are smaller in the former than in the latter case. This bit rate increase in intraframe coding causes the proposed algorithm to provide a larger PSNR improvement in I than in P frames as indicated in Table I of the following section. Typical values of the GOP period (10–15 frames) in videoconference applications

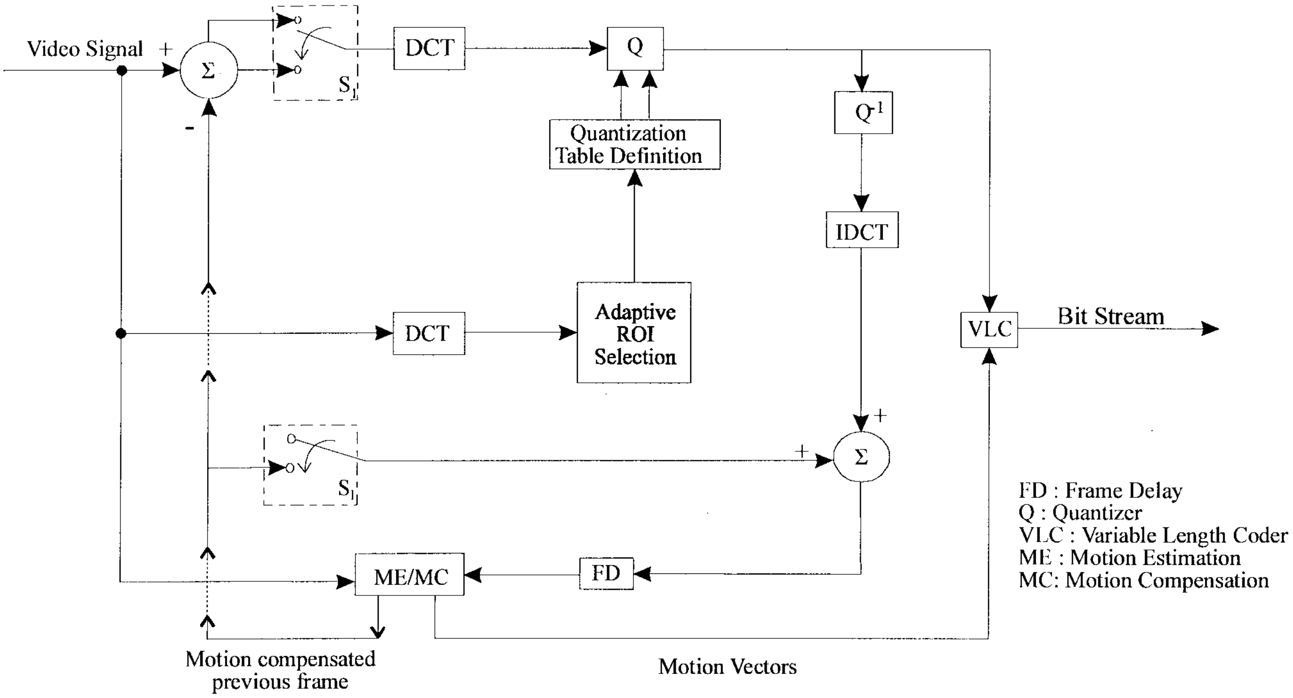


Fig. 2. ROI-based MC–DCT encoder.

ensure that there is no accumulation of the motion-estimated prediction error.

Fig. 2 presents the proposed ROI-based MC–DCT coding scheme. Apart from the conventional parts of MPEG-1 encoders, such as quantization, motion estimation/compensation, and entropy coding, a foreground VOP (ROI) selection unit has been added, regulating the operation of the rate control mechanism. This unit operates in the DCT domain. When the S_1 switch activates intraframe coding, a delay occurs which permits the neural network architecture to perform the frame segmentation task. When the S_1 switch activates interframe coding, the DCT coefficients are computed before the network operation since they are not available directly from the bit stream. Using hardware implementations of the fast DCT and the LVQ algorithm, real-time system operation is feasible.

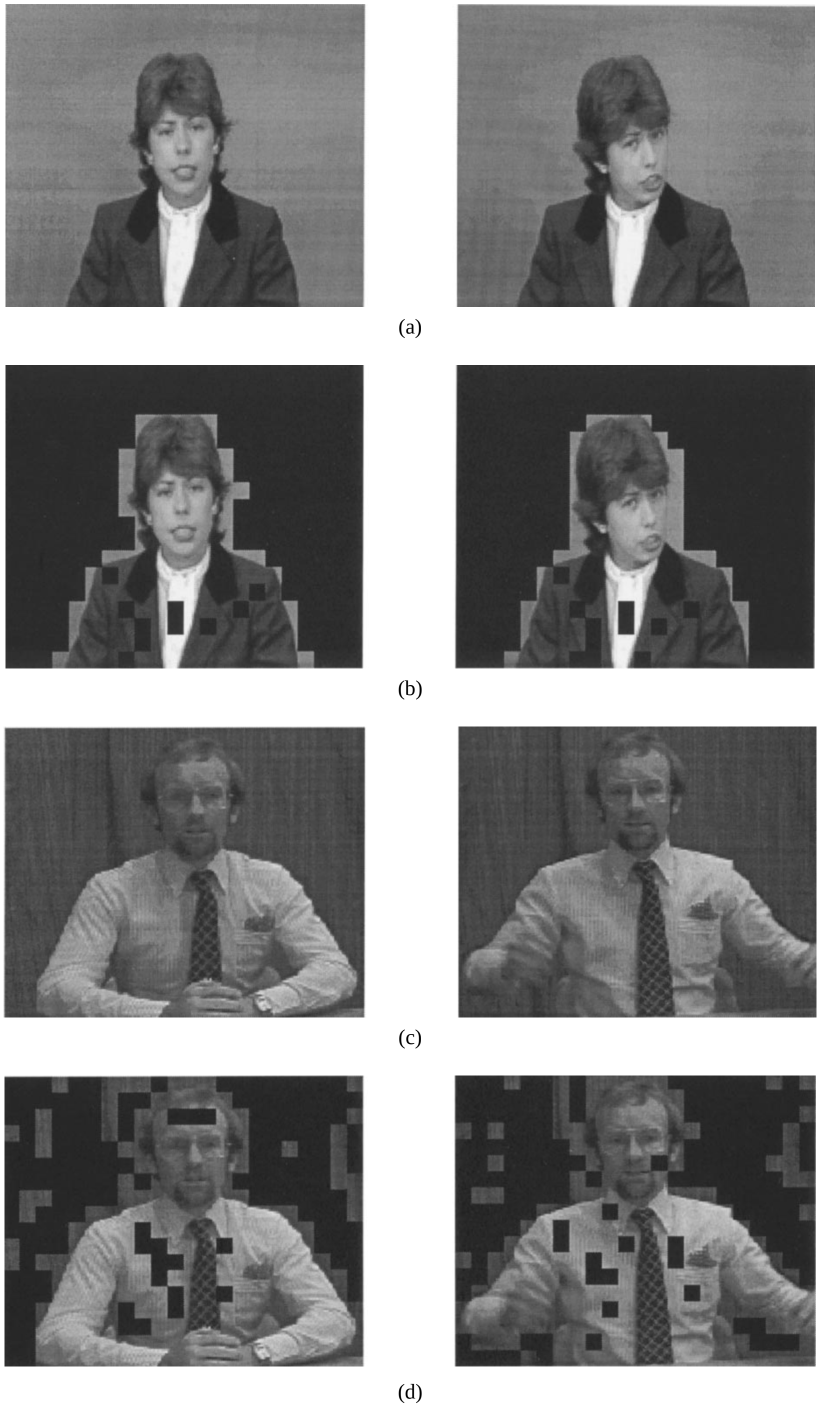
IV. Simulation results

The performance of the proposed neural-network-based system was evaluated using the Claire and Trevor image sequences. These were in QCIF format, with all components having the same size (176 × 144 pixels). A picture rate of 10 frames/s was used.

The first 24 (3 × 8) zig-zag scanned ac DCT coefficients of the three color components of each block were used as inputs features to the first-level neural network classifier. The second network classifier was fed with the above coefficients as well as with the corresponding dc coefficients (27 inputs). The former network was trained using all blocks of the first and second frames of the Claire sequence. Its performance was tested using the remaining 148 frames of the same sequence, as well as all frames of the Trevor sequence. The second network was also initially trained with the first two frames of Claire, and was automatically retrained when detecting the change of the operational environment that occurred during transition from Claire to Trevor.

Fig. 3(a) presents frames 30 and 100 of Claire which are different from the ones used for training the network. Fig. 3(b) shows the corresponding binary segmentation masks provided by the first network. For clarity of presentation, when a block belongs to background, the values of its pixels are set to zero (black pixels); foreground blocks are shown as they are. Fig. 3(c) shows two frames of Trevor with arms in closed and open position, while Fig. 3(d) shows the respective segmentation masks provided by the first network. It can be easily seen that the first network generalizes well, providing segmentation masks of good quality. There are, however, some misclassified blocks; in the case of Claire, where the background is rather uniform, such misclassifications occur in the foreground VOP which contains some homogeneous regions, such as clothes and forehead. In the case of Trevor, where background is less uniform, it contains some misclassifications as well. Using, however, additional color information within a scene, and by applying the scene detection and network retraining mechanisms described in Section II, the second neural network was able to correct these misclassifications, providing the images shown in Fig. 4.

The philosophy of the proposed approach differs from that of conventional segmentation algorithms. The latter try to find spatially or temporally uniform, large or small, segments of the images [5], [6]. These segments may not, however, correspond to physical objects in the scene; consequently, pre- or postprocessing, using semiautomatic techniques, should be combined with such methods to lead to physical object extraction. In the current approach, a highly nonlinear separation of the feature space is performed by the neural network classifiers which can also account for dynamic changes of the environment.



(a)

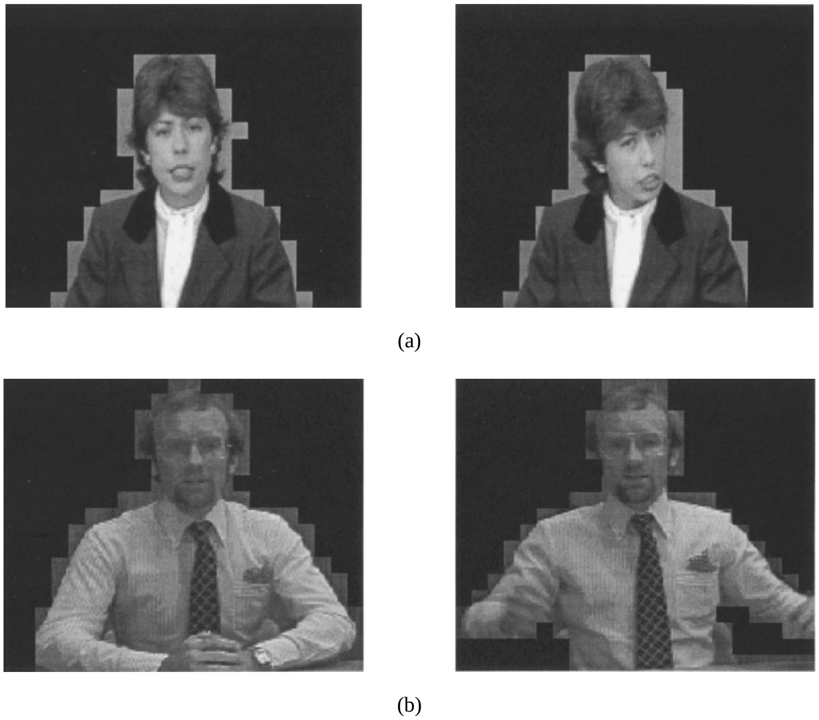
(b)

(c)

(d)

Fig. 3. (a) Original frames 30 and 100 of Claire. (b) Corresponding segmentation masks provided by the first neural network. (c) Original frames 35 and 130 of Trevor. (d) Corresponding segmentation masks provided by the first neural network.

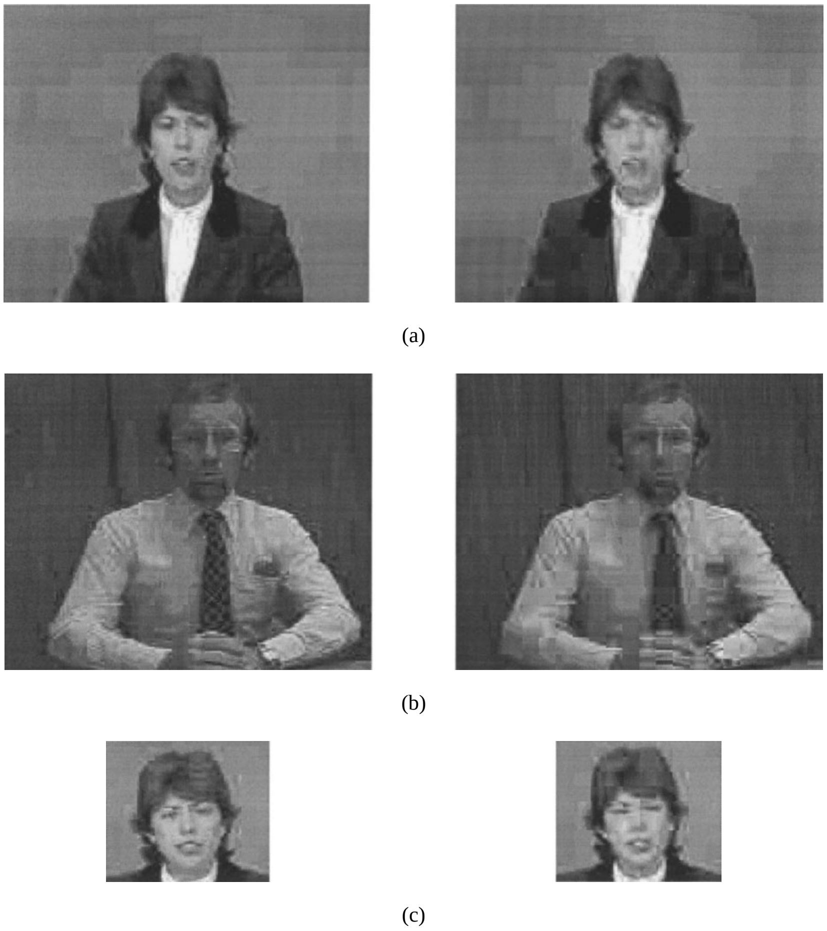
In the following, we present a comparison between the proposed ROI-based MC–DCT and the conventional MPEG-1 algorithm. Table I shows the average peak SNR (PSNR) obtained by the two coding schemes for intraframe and interframe coding of the Claire and Trevor sequences, using 150 frames from each sequence. An average improvement of PSNR about 1.4 dB has been observed. P frames have been only used for interframe coding, with an intraframe distance of 10. PSNR improvement was larger in case of I than P frames while, on average, it was close to that of P frames, which were the majority within each group of pictures. The last two columns of Table I present the proportion of bits allocated to ROI blocks in intra- as well as in interframe coding. The proportion is smaller in Claire due to the fact that ROI areas occupy a smaller part of the image. As the bit rate reduces, the percentage of ROI to total bits reduces as well since the high-frequency content of foreground is eliminated by the algorithm so as to achieve the low bit rate, while background regions are almost saturated. In interframe coding, the percentage of bits allocated to ROI is much smaller due to the fact that the major part of available bits is allocated to coding of motion vectors.



(a)

(b)

Fig. 4. (a) Segmentation masks of frames 30 and 100 of Claire provided by the second neural network after retraining with the additional features. (b) Segmentation masks of frames 35 and 130 of Trevor provided by the second network after automatic retraining with frame 1 of Trevor.



(a)

(b)

(c)

Fig. 5. Reconstructed frames. (a) 21st frame of Claire at 16 kbit/s using ROI MC–DCT and MC–DCT algorithms. (b) Forty-first frame of Trevor at 40 kbit/s using ROI MC–DCT and MC–DCT algorithms. (c) Zooming at the facial area of (a).

Fig. 5(a) presents the decoded images of Claire provided by the proposed approach, as well as by the MPEG-1 algorithm, at 16 kbit/s. The quality improvement provided by the former technique can be easily discerned. Similar results hold for Trevor [Fig. 5(b)], where degradation of the quality of background is more visible. A zooming on the head part of Claire at 16 kbit/s is performed in Fig. 5(c), showing the good quality achieved by the proposed approach in foreground VOP's.

V. Conclusions and Further Work

An adaptive technique for extracting VOP's in image sequences has been proposed in this paper. A neural network subsystem has been designed for selecting foreground VOP's (ROI), based on the information included in the DCT coefficients of each transformed block of the images. This scheme has been implemented within an MPEG-1 framework, providing improvement of PSNR as well as reconstructed images of good quality.

The proposed technique can be similarly incorporated, either in H.263 or in the forthcoming MPEG-4 standards, in which each VOP may be characterized by different frame rate, resolution, and quality, and in which a stationary background may be transmitted, not at every frame, but only once at the beginning of each scene. The focus of the paper has been on videophone and videoconference applications. Video surveillance, image data base browsing [9], and medical image compression and transmission [10] are other applications which can take advantage of the proposed approach.

Examples have been presented which illustrate the performance of the method when dealing with almost uniform background. Based on its training capabilities, the system can learn to classify specific types of nonuniform background to the non-ROI categories. We are currently working on extensions of the neural network system to effectively handle such cases as well.

References

[1] L. Chiariglione, "MPEG and multimedia communications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 5–18, Feb. 1997.

[2] ITU-T SG 15 Experts Group for Very Low Bit Rate Visual Telephony, Draft Recommendation H.263, Feb. 1995.

[3] T. Sikora, "The MPEG-4 video standard verification model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 19–31, Feb. 1997.

[4] MPEG Video Group, "MPEG-4 Video Verification Model—Version 5.0," Doc. ISO/IEC/JTCI/SC29/WG11, N1469 Maceio, Nov. 1996.

[5] P. Salembier and M. Pardas, "Hierarchical morphological segmentation for image sequence coding," *IEEE Trans. Image Processing*, vol. 3, pp. 639–651, Sept. 1994.

[6] F. Meyer and S. Beucher, "Morphological segmentation," *J. Visual Commun. Image Representation*, vol. 1, pp. 21–46, Sept. 1990.

[7] E. Reusens, T. Ebrahimi, C. Le Buhan, R. Castagno, V. Vaerman, C. de Sola Fabregas, S. Bhattacharjee, F. Bossen, and M. Kunt, "Dynamic approach to visual data compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 197–211, Feb. 1997.

[8] S. Haykin, *Neural Networks: A Comprehensive Foundation.* New York: Macmillan, 1994.

[9] A. Doulamis, Y. Avrithis, N. Doulamis, and S. Kollias, "Indexing and retrieval of the most characteristic frames/scenes in video databases," in *Proc. Workshop Image Analysis for Multimedia Interactive Services (WIAMIS'97)*, Louvain-la-Neuve, Belgium, June 1997, pp. 105–110.

[10] N. Panagiotidis, D. Kalogeras, S. Kollias, and A. Stafylopatis, "Neural network assisted effective lossy compression of medical images," *Proc. IEEE*, vol. 84, pp. 1474–1487, Oct. 1996.