An Efficient Fully Unsupervised Video Object Segmentation Scheme Using an Adaptive Neural-Network Classifier Architecture

Anastasios Doulamis, Nikolaos Doulamis, Klimis Ntalianis, and Stefanos Kollias

***Abstract*—In this paper, an unsupervised video object (VO) segmentation and tracking algorithm is proposed based on an adaptable neural-network architecture. The proposed scheme comprises: 1) a VO tracking module and 2) an initial VO estimation module. Object tracking is handled as a classification problem and implemented through an adaptive network classifier, which provides better results compared to conventional motion-based tracking algorithms. Network adaptation is accomplished through an efficient and cost effective weight updating algorithm, providing a minimum degradation of the previous network knowledge and taking into account the current content conditions. A retraining set is constructed and used for this purpose based on initial VO estimation results. Two different scenarios are investigated. The first concerns extraction of human entities in video conferencing applications, while the second exploits depth information to identify generic VOs in stereoscopic video sequences. Human face/body detection based on Gaussian distributions is accomplished in the first scenario, while segmentation fusion is obtained using color and depth information in the second scenario. A decision mechanism is also incorporated to detect time instances for weight updating. Experimental results and comparisons indicate the good performance of the proposed scheme even in sequences with complicated content (object bending, occlusion).**

***Index Terms*—Adaptive neural networks, MPEG-4, video object extraction.**

I. Introduction

NEW multimedia applications, such as video editing, content-based image retrieval, video summarization, object-based transmission and video surveillance strongly depend on characterization of the visual content. For this reason, the MPEG-4 coding standard has introduced the concepts of video objects (VOs), which correspond to meaningful (semantic) content entities, such as buildings, ships or persons. VOs consist of regions of arbitrary shape with different color, texture, and motion properties. Such object-based representations offer a new range of capabilities in terms of accessing and manipulating visual information [1]. In particular, *high compression ratios* are achieved by allowing the encoder to place more emphasis on objects of interest [1], [2]. Furthermore, sophisticated *video queries* and *content-based retrieval operations* on image/video databases are effectively performed [3], [4]. Although the MPEG-4 standard has introduced the concept of VOs and specified a general coding methodology and a syntax for them, it left the problem of object extraction to content developers. VO extraction remains a very interesting and challenging task.

Color segmentation can be considered as a first step toward VO extraction. Some typical works include the morphological watershed [5], the split and merge technique [6] or the recursive shortest spanning tree (RSST) algorithm [7]. However, an intrinsic property of VOs is that they usually consist of regions of totally different color characteristics. Consequently the main problem of any color-oriented segmentation scheme is that it oversegments an object into multiple color regions. A more meaningful content representation is provided by exploiting motion information [8]–[10]. However, in this case, object boundaries cannot be identified with high accuracy mainly due to erroneous estimation of motion vectors. For this reason, hybrid approaches have been proposed which efficiently combine color/motion properties [11], [12]. Although these methods provide satisfactory results for VOs of homogeneous motion characteristics, they fail in detecting objects of no coherent motion. This is, for example, the case of a news-speaker who moves only his/her head or hands, while keeps his/her body still.

More accurate VO extraction is achieved by considering the user as a part of the segmentation process, resulting in semiautomatic segmentation schemes. In this framework, the user initially provides a coarse approximation of the VO and then the system performs object segmentation by refining the initial user's approximation [13], [14]. Semiautomatic segmentation schemes can be also considered VO tracking algorithms [15]–[17]. These algorithms dynamically track object contours through time based on an initial approximate of the VO. Most of tracking algorithms update object boundaries by exploiting motion information. Although such approaches provide good results for slow varying object boundaries, they are not suitable in complex situations, where high motion or abrupt variations of object contours are encountered.

On the other hand, neural networks can become major image/video analysis tools due to their highly nonlinear capabilities. However, before applying a neural network to real-life applications two main issues should be effectively confronted. The first concerns network generalization. Many significant results have been derived toward this direction during the last few years [18]. Examples include algorithms for adaptive creation of the network architecture during training, such as pruning or constructive techniques [19], [20], or theoretical

Manuscript received January 3, 2003. This work was supported in part by the National Scholarship Foundation.

The authors are with the Electrical and Computer Engineering Department, National Technical University of Athens, Zografou 15773, Athens, Greece (e-mail: adoulam@cs.ntua.gr).

Digital Object Identifier 10.1109/TNN.2003.810605

aspects, such as the VC dimension [21]. Specific results and mathematical formulations regarding error bounds and overtraining issues have been obtained when considering cases with known probability distributions of the data or stationarity of the operational environment [22], [23]. Despite, however, the achievements obtained, in most cases, VOs do not obey some specific probability distribution. That is why direct application of conventional neural networks is not always adequate for solving object detection and classification problems. The second issue originates from the high computational cost usually required for network training, which becomes crucial when dealing with real-time applications, such as video analysis problems.

To overcome the aforementioned difficulties, an adaptable neural-network architecture is used in this paper, based on an efficient weight updating algorithm which adapts network performance to current conditions. Network adaptation improves network generalization, since it exploits information about current conditions. In particular, network weights are updated so that: 1) a minimum degradation of the previous network knowledge is provided, while 2) the current conditions are trusted as much as possible. Furthermore, the proposed weight updating algorithm is implemented in a cost-effective manner, so that the adaptable architecture can be applied to real-time applications.

Taking the previously mentioned issues into consideration, we propose an unsupervised VO segmentation and tracking scheme that incorporates an adaptable neural-network architecture. The scheme is investigated into two interesting different scenarios. The first concerns extraction of human entities in video conferencing sequences. The second exploits depth information, provided by a multiview camera system (e.g., stereoscopic sequences), to perform extraction of generic VOs. A general overview of the proposed system is presented in Fig. 1. As can be seen, the system comprises two different modules: 1) the initial VO segmentation module, which provides a coarse object approximation and 2) the object tracking module, which is accomplished by the adaptable neural-network architecture. Weight updating is based on a retraining set that is constructed to describe the content of the current environment. Finally, a decision mechanism is also incorporated to activate network retraining at time instances of significant visual content variations.

In the first scenario, initial VO estimation is achieved by a human face and body detector. A human face is localized based on a Gausssian probability density function (pdf) that models the color components of the facial area and a shape constraint about the estimated color-face regions, using a binary template matching technique. Human body detection relies on a probabilistic model, the parameters of which are estimated according to the center, height, and width of the extracted facial region. In the second scenario, depth information is exploited for initial VO estimation. Depth information is an important feature for semantic segmentation, since usually VOs are located on the same depth plane. A two-camera system (stereoscopic) is considered to reliably estimate depth information.

The main contribution of this paper is focused on the following issues. An unsupervised content-based segmentation and tracking scheme of semantically meaningful VOs is presented in a cost effective and efficient manner. Two different scenarios are investigated. The first refers to the extraction of human entities in videophone/video conferencing applications, while the second to the identification of generic video entities by exploiting depth information in video sequences captured with two (or) more cameras (e.g., stereoscopic sequences). The unsupervised VO extraction scheme includes two main tasks; the initial VO estimation and the accurate tracking of VOs through time. In this paper, we contribute on each of both aspects as well as on their integration.

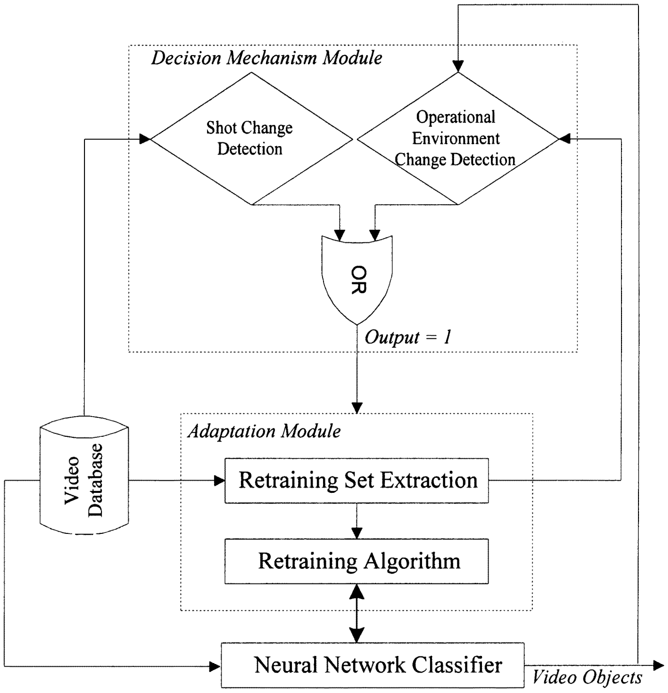

Decision Mechanism Module
Shot Change Detection
Operational Environment Change Detection
OR
Output = 1
Adaptation Module
Video Database
Retraining Set Extraction
Retraining Algorithm
Neural Network Classifier
Video Objects

Fig. 1. Overview of the proposed adaptive neural-network classifier architecture.

VO tracking is handled, in our case, as a nonlinear classification problem; each VO as a class defines a class assigning each examined image region is assigned to one of the available classes. This consideration provides more accurate results compared to conventional algorithms, in which VOs are tracked by exploiting motion information; this is more evident in sequences with complicated content, where high motion, complex background, special camera effects (zooming, panning) or occluded regions are encountered. VO tracking is performed based on the adaptable neural-network architecture. This is an often case in real-life video sequences due to variations of luminosity conditions or of shot content. Instead, conventional neural networks, where network weights are considered constant through time, are not potentially useful for such dynamic environments. Furthermore, the proposed weight updating algorithm is implemented in a cost effective way, permitting video tracking in real time, while simultaneously retains the tracking efficiency, since the algorithm guarantees that its performance to the current data is satisfactory.

As far as the initial VO estimation is concerned, a novel color based modeling of human faces is proposed based on Gausssian probability distributions, while human body is probabilistically detected with respect to the human face location. The aforementioned probabilistic scheme is applied to the scenario dealing with video conferencing applications, while for the stereo sequences, the initial VO estimation is provided through a constrained color/depth segmentation algorithm. In both cases, only a coarse approximate of the initial VO mask is required, since accurate object segmentation and tracking is performed through the adaptable neural-network architecture. Therefore, the system performance is enhanced since it is more easily and less computationally demanded to provide a coarse approximation of an object than an accurate one. Furthermore, a decision mechanism is also incorporated, which evaluates the network performance and estimates the time instances of weight updating avoiding user interaction.

Theoretical aspects concerning network retraining have been investigated in [24]. However, in [24], we do not concentrate on content-based segmentation and tracking in real-life sequences of complicated content. Instead, the theoretical results are evaluated for image analysis problems and using images of simple content (uniform background). Furthermore, in [25], depth information is combined with color characteristics for object segmentation without examining the cases of VO tracking. Furthermore, in this paper, no neural networks have been applied.

The paper is organized as follows: Section II presents the proposed neural-network retraining strategy. In Section III, we investigate initial VO estimation, in the first scenario trhough human face and body modeling. In Section IV, construction of the retraining set is examined, in the second scenario, exploiting depth information. Specifications of the decision mechanism are deployed in Section V. Experimental results and comparisons with known techniques are presented in Section VI. Finally, Section VII concludes the paper.

II. Neural–Network Retraining Strategy

A. Formulation of the VO Extraction Problem (VOP)

VO extraction is treated next as a classification problem. This means that each image pixel p_i is assigned to one of, say, M available classes ω_i, $i = 1, 2, \ldots, M$, each of which corresponds to a particular VO. To classify a pixel p_i, a feature vector $\mathbf{b}_i$ is estimated on a block of 8×8 pixels centered around the examined pixel $\mathbf{b}_i$. Let us assume that a neural network is used to perform the classification task, i.e., the VO extraction. Then, the neural-network output is written as

$$\mathbf{y}(\mathbf{b}_i) = \left[p^i_{\omega_1} p^i_{\omega_2} \ldots p^i_{\omega_M}\right]^T \tag{1}$$

where $p^i_{\omega_j}$ refers to the degree of coherence of $\mathbf{b}_i$ to class ω_j.

Consider now that the neural network has been initially trained using a set of m_b pairs of the form $S_b = \{(\mathbf{b'}_1, \mathbf{d'}_1), \ldots, (\mathbf{b'}_{m_b}, \mathbf{d'}_{m_b})\}$, where $\mathbf{b'}_i$ and $\mathbf{d'}_i$, with $i = 1, 2, \ldots, m_b$, denote the ith input training vector and the corresponding desired output vector. Vectors $\mathbf{b'}_i$ and $\mathbf{d'}_i$ have the same form as vectors $\mathbf{b}_i$ and $\mathbf{y}(\cdot)$, respectively.

However, in a VO extraction problem several changes of the visual operational environment occur and thus, the weights of the network classifier cannot be considered constant. This is, for example, the case of a video sequence consisting of several scenes of different VO properties. For this reason, the network weights should be adapted resulting in an adaptable neural-network architecture. In particular, network weights are modified through a retraining set, denoted as S_c, comprising m_c pairs of the form $S_c = \{(\mathbf{b}_1, \mathbf{d}_1), \ldots (\mathbf{b}_{m_c}, \mathbf{d}_{m_c})\}$, where $\mathbf{b}_i$ and $\mathbf{d}_i$ with $i = 1, 2, \ldots, m_c$ correspond to the ith input vector and the desired network output, respectively.

Weight updating is performed to minimize the error over the samples of set S_b and set S_c

$$E_a = E_{c,a} + \eta E_{f,a} \tag{2}$$

with

$$\begin{aligned} E_{c,a} =& \frac{1}{2} \sum_{i=1}^{m_c} \left\| \mathbf{z}_a\left(\mathbf{b}_i\right) - \mathbf{d}_i \right\|_2 \\ E_{f,a} =& \frac{1}{2} \sum_{i=1}^{m_b} \left\| \mathbf{z}_a\left(\mathbf{b'}_i\right) - \mathbf{d'}_i \right\|_2 \end{aligned} \tag{2a}$$

where $E_{c,a}$ is the error over training set S_c, $E_{f,a}$ the corresponding error over the training set S_b; $\mathbf{z}_a(\mathbf{b}_i)$ and $\mathbf{z}_a(\mathbf{b'}_i)$ are the outputs of the network *after* retraining, corresponding to input vectors $\mathbf{b}_i$ and $\mathbf{b'}_i$, respectively. Similarly $\mathbf{z}_b(\mathbf{b}_i)$ is the network output *before* retraining. Parameter η is a weighting factor, accounting for the significance of the current training set compared to the former one.

B. Network Retraining Algorithm

Let us, for simplicity, consider: 1) a two-class classification problem, where classes ω_1, ω_2 refer to the foreground and background VO of an image and 2) a feedforward neural-network classifier, which includes a single output neuron, one hidden layer consisting of q neurons, and an input layer of J elements (J is the size of feature vector $\mathbf{b}_i$ or $\mathbf{b}'_i$). Then, the network output with input is expressed as [18]

$$z_{\{a,b\}}\left(\mathbf{b}_i\right) = f\left(\left(\mathbf{w}^1_{\{a,b\}}\right)^T \cdot \mathbf{u}_{\{a,b\}}\left(\mathbf{b}_i\right)\right)$$

with

$$\mathbf{u}_{\{a,b\}}\left(\mathbf{b}_i\right) = \mathbf{f}\left(\left(\mathbf{W}^0_{\{a,b\}}\right)^T \cdot \mathbf{b}_i\right) \tag{3}$$

where subscripts $\{a, b\}$ refer to the states *"after"* or *"before"* retraining. The $\mathbf{w}^1_{\{a,b\}}$ denotes a vector containing the $q \times 1$ weights between the output and hidden neurons and $\mathbf{W}^0_{\{a,b\}}$ is a $J \times q$ matrix defined as $\mathbf{W}^0_{\{a,b\}} = [\mathbf{w}^0_{1,\{a,b\}}, \ldots, \mathbf{w}^0_{q,\{a,b\}}]$, where $\mathbf{w}^0_{k,\{a,b\}}$, $k = 1, 2, \ldots, q$, corresponds to a $J \times 1$ vector of the weights between the kth hidden neuron and the network inputs. $f(\cdot)$ is the sigmoid function, while vector $\mathbf{f}(\cdot)$ is a vector-valued function with elements the functions $f(\cdot)$. In (3), the network output $z_{\{a,b\}}(\mathbf{b}_i)$ is a scalar, since we have considered a two-class classification problem. Values of $z_{\{a,b\}}(\mathbf{b}_i)$ close to zero indicate the backgound object, while values close to one the foreground object.

The goal of the retraining procedure is to estimate the weights after retraining $\mathbf{w}_a$, i.e., $\mathbf{W}^0_a$ and $\mathbf{w}^1_a$, respectively. Assuming that a small perturbation of the weights is sufficient to keep a reliable performance to the new environment, we conclude that

$$\mathbf{W}_a^0 = \mathbf{W}_b^0 + \Delta\mathbf{W}^0 \mathbf{w}_a^1 = \mathbf{w}_b^1 + \Delta\mathbf{w}^1 \tag{4}$$

where $\Delta\mathbf{W}^0$ and $\Delta\mathbf{w}^1$ are small weight increments. Furthermore to stress the importance of current data, one can replace the first term of (2a) by the constraint that the actual network outputs are equal to the desired ones, that is

$$z_a(\mathbf{b}_i) = d_i \quad i = 1, \ldots, m_c, \qquad \text{for all data in } S_c \tag{5}$$

where in this case d_i is scalar, since we refer to a 2-class classification problem.

Since (4) is held, the neuron activation functions can be linearized and, thus, (5) is equivalent to a set of linear equations

$$\mathbf{c} = \mathbf{A} \cdot \Delta\mathbf{w} \tag{6}$$

where $\Delta\mathbf{w}$ is the vector containing the small perturbation of all network weights. Vector $\mathbf{c}$ and matrix A are appropriately expressed in terms of the previous network weights [24].

The number of linear equations in (6) is in general smaller than the number of unknown weights $\Delta\mathbf{w}$, since a small number of retraining data m_c is usually chosen. Therefore, uniqueness is imposed by the additional requirement of minimum degradation of the previous network knowledge, i.e.,

$$E_S = \|E_{f,a} - E_{f,b}\|_2 \tag{7}$$

with $E_{f,a}$ defined similarly to $E_{f,a}$ when $\mathbf{z}_a$ is replaced by $\mathbf{z}_b$ in the right-hand side of the second term in (2a). It has been shown in [24] that (7) takes the form

$$E_S = \frac{1}{2}(\Delta\mathbf{w})^T \cdot \mathbf{K}^T \cdot \mathbf{K} \cdot \Delta\mathbf{w} \tag{8}$$

where the elements of matrix $\mathbf{K}$ are expressed in terms of the previous network weights $\mathbf{w}_b$ and the training data in S_b. Thus, the problem results in minimization of (8) subject to the linear constraints of (6) and a solution is estimated by adopting the gradient projection method [26].

III. Unsupervised Retraining Set Estimation in Videoconference Sequences

The aforementioned weight updating is performed each time the visual content changes. These variations are detected by the decision mechanism, which activates a new retraining phase. Perhaps the most important issue of network retraining is to efficiently estimate the content characteristics of the current environment, which is equivalent of estimating a reliable retraining set S_c. The retraining set is constructed through the initial VO estimation module. This is investigated in this section, where human entities are extracted from background in video conferencing applications (first scenario).

Initial estimation of human entities is provided through: 1) a human face and 2) a human body detection module. In particular, human faces are first detected using a color-based model, followed by a template matching algorithm, while human bodies are localized based on a probabilistic model, the parameters of which are estimated according to the detected facial region. In the following, the techniques for human face and body detection are analytically described.

A. Human Face Detection

Various methods and algorithms have been proposed in the literature over the years, for human face detection, ranging from edge map projections to recent techniques using generalized symmetric operators [27], [28]. In the proposed approach, the two-chrominance components of a color image are used for efficiently performing human face detection, as the distribution of the chrominance values of a human face, occupies a very small region of the color space [29]. Thus, blocks whose respective chrominance values are located in this small region, can be considered as facial blocks. On the contrary, blocks of chrominance with values located far from this region correspond to nonface blocks. The computational complexity of this method is significantly low, which is very important property, especially in case of videoconference applications.

The histogram of chrominance values corresponding to face class Ω_f is initially modeled by a Gausssian probability density function (pdf), instead of using a Bayessian approach as in [29]. As a result

$$P(\mathbf{x} \mid \Omega_f) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \mu_f)^T \cdot \mathbf{\Sigma}_f^{-1} \cdot (\mathbf{x} - \mu_f)\right)}{2\pi \cdot |\mathbf{\Sigma}|^{1/2}} \tag{9}$$

where $x = [uv]^T$ is a 2×1 vector containing the mean chrominance components u and v of an examined image region. Since only an approximation of the human face is adequate to perform the initial VO estimation, image regions corresponding to nonoverlapping blocks of 8×8 pixels are used for face detection. Vectors μ_f matrix $\mathbf{\Sigma}$ of (9) express the mean and variance of the chrominance values over a training set. In our implementation, a confidence interval of 80% is selected so that only blocks inside this region are considered as face blocks, while blocks belonging to the rest 20% as nonface blocks.

However, as the aforementioned procedure takes into consideration only color information, the estimated mask may also contain nonface blocks, which present similar chrominance properties to face regions, e.g., human hands. To confront this difficulty, the shape information of the human face is also taken into consideration. More specifically, the shape of human faces is unique and consistent; its boundary can be approximated by an ellipse, by connected arcs or by a rectangle [29]. The method of [29] is adopted next, where rectangles of certain aspect ratios are used for the approximation of the human face. In particular, in the adopted scheme, the rectangle aspect ratio is defined as

$$R = \frac{H_f}{W_f} \tag{10}$$

where H_f is the height of the head, while W_f corresponds to the face width. Using several experiments, R was found to lie within the interval [1.2 1.7]. Furthermore, blocks of size smaller than 8×8 pixels are ignored, since we consider that they correspond to noise.

B. Human Body Detection

Follow ing human face detection, the human body is localized by exploiting information derived from the human face detection module. In particular, initially the center, width and height of the face region, denoted by $\mathbf{c}_f = [c_x c_y]^T$, w_f and h_f, respectively, are calculated. The human body is then localized by incorporating a probabilistic model, the parameters of which are estimated according to the $\mathbf{c}_f$, w_f and h_f parameters.

Again, the following analysis is concentrated on a block resolution since we are interested only in an approximation of the human body. In particular, let us denote by $\mathbf{r}(B_i) = [r_x(B_i) r_y(B_i)]^T$ denoting the distance between the ith block, B_i, from the origins, with $r_x(B_i)$ and $r_y(B_i)$ are the respective x and y coordinates. The product of two independent one-dimensional Gaussian pdfs is used next for body modeling. Particularly, for each block B_i of an image, a probability $P(\mathbf{r}(B_i)|\Omega_b)$ is assigned, expressing the degree of block B_i belonging to the human body class Ω_b.

$$\begin{aligned} &P\left(\mathbf{r}\left(B_i\right)|\Omega_b\right) \\ &= \frac{\exp\left(-\frac{1}{2\sigma_x^2}\left(r_x\left(B_i\right)-\mu_x\right)^2\right)\exp\left(-\frac{1}{2\sigma_y}\left(r_y\left(B_i\right)-\mu_y\right)^2\right)}{(2\pi)\sigma_x\sigma_y} \end{aligned} \tag{11}$$

where $\mu_x \mu_y$, σ_x, and σ_y express the parameters of the human body location model; these parameters are calculated based on information derived from the face detection task, taking into account the relationship between human face and body

$$\mu_x = c_x,\ \mu_y = c_y + h_f \tag{12}$$

$$\sigma_x = w_f \sigma_y = \frac{h_f}{2}. \tag{13}$$

A confidence interval of 80% is selected from the Gausssian model so that blocks belonging to this interval are considered as human body blocks.

C. Retraining Set Construction

The human face and body detection modules provide an initial estimate of the foreground object. This information is then used to construct the retraining set S_c, which describes the content of the current environment. In particular, all blocks that have been classified either to a face or a body region are included in set S_c as foreground data. The remaining blocks, however, cannot be included in the retraining set S_c as background data, since the initial VO estimation module is only an approximation of the human entity and thus, selecting data blocks adjacent to the foreground object as background may confuse the network during retraining. For this reason, a region of uncertainty is created around the selected foreground samples (human face and body) and the blocks of it are excluded from the retraining set S_c. The uncertainty region is formed by creating a zone of blocks around the detected human face and body region, regulating, for example, the confidence interval of the probabilistic model.

IV. Unsupervised Retraining Set Extraction by Exploiting Depth Information

In this section, we deal with the second scenario, in which VO segmentation is performed by exploiting depth information. Depth is an important element toward content description since, a VO is usually located on the same depth plane [25]. Depth is reliably estimated by using stereoscopic (or multiview) imaging techniques and this case is examined in the following.

However, VO boundaries cannot be identified with high accuracy by a depth segmentation algorithm, mainly due to erroneous estimation of the disparity field and occlusion issues. To confront this difficulty, color information is exploited, which retains reliable object boundaries, but usually oversegments VOs into multiple regions. In particular, color segments are fused with depth segments to provide the initial estimate of the VO.

A. Color and Depth Information Fusion

Let us assume that K^c color segments and K^d depth segments have been extracted by applying a segmentation algorithm to the color-depth information of an image respectively. Color (depth) segments are denoted as $F_i^c(F_i^d)$, $i = 1, 2, \ldots, K^c(K^d)$. Let us also denote by G^c and G^d the output masks of color and depth segmentation, which are defined as the sets of all color and depth segments, respectively

$$\begin{aligned} G^c &= \{F_i^c, i = 1, 2, \ldots, K^c\} \\ G^d &= \left\{F_i^d, i = 1, 2, \ldots, K^d\right\}. \end{aligned} \tag{14}$$

Color segments are projected onto depth segments so that VOs provided by the depth segmentation are retained and, at the same time, object boundaries given by the color segmentation are accurately extracted. For this reason, each color segment F_i^c is associated with a depth segment, so that the area of intersection between the two segments is maximized. This is accomplished by means of a *projection function*

$$p\left(F_i^c, G^d\right) = \arg\max_{g \in G^d} \left\{a\left(g \cap F_i^c\right)\right\}, \qquad i = 1, 2, \ldots, K^c \tag{15}$$

where $a(\cdot)$ returns the area, i.e., the number of pixels, of a segment. Based on the previous equation, K^d sets of color segments, say C_i, $i = 1, 2, \ldots, K^d$, are created, each of which contains all color segments that are projected onto the same depth segment F_i^d

$$C_i = \left\{g \in G^c : p\left(g, G^d\right) = F_i^d\right\}, \qquad i = 1, 2, \ldots, K^d. \tag{16}$$

Then, the final segmentation mask, G, consists of $K = K^d$ segments F_i, $i = 1, 2, \ldots, K$, each of which is generated as the union of all elements of the corresponding set C_i

$$F_i = \bigcup_{g \in C_i} g, \qquad i = 1, 2, \ldots, K \tag{17}$$

$$G = \{F_i, \qquad i = 1, 2, \ldots, K\}. \tag{18}$$

In other words, color segments are merged together into $K = K^d$ new segments according to depth similarity.

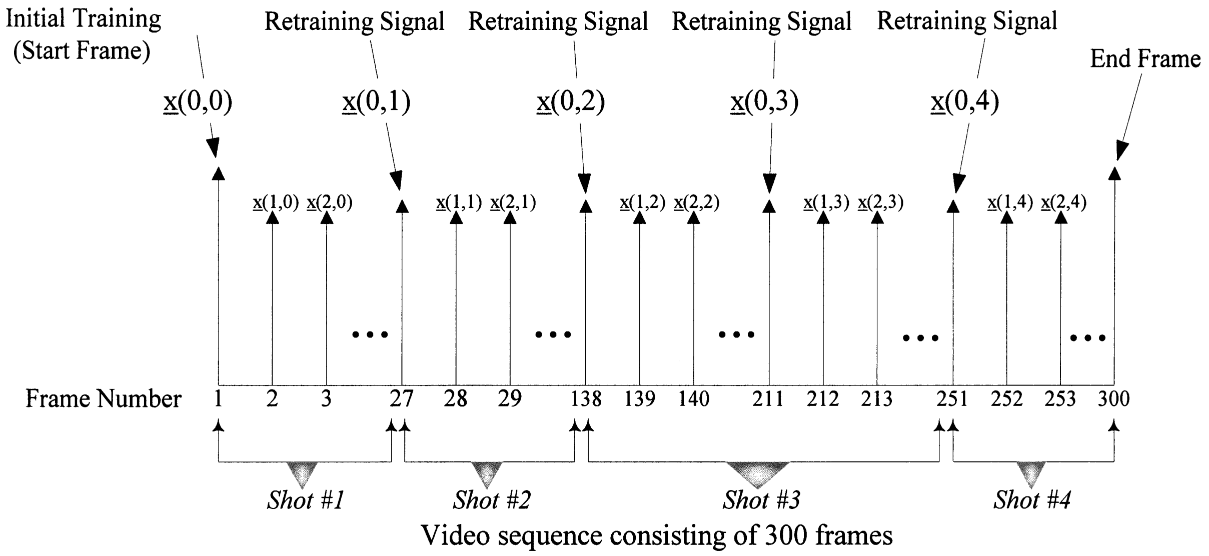


Fig. 2. Four retraining time instances of a sequence of four shots and 300 frames. Two retraining signals are generated for shot #3.

V. Specifications of the Decision Mechanism

The decision mechanism plays a very important role in the proposed scheme, since each time it generates an alarm signal, the retraining module is activated. Several alarms may significantly increase the computational complexity of the proposed scheme, mainly due to the initial video estimation module (retraining set construction) and the weight updating algorithm. On the other hand, a small number of alarms may induce serious deterioration of the tracking performance. Thus, when implementing the decision mechanism, the tradeoff between reliable network performance and minimal cost of the procedure should be effectively considered.

In the proposed approach, the decision mechanism consists of a shot cut detection module and an operational environment change module. The first is based on the principle that all different poses that a VO takes within a shot are usually strongly correlated to each other, while the second is incorporated as a safety valve to confront gradually but significantly content changes within a shot. The first module permits parallelization of the process, since different shots can be processed independently.

An example of alarm signals provided by the decision mechanism is depicted in Fig. 2. More specifically, let us index images or video frames in time, denoting by $\mathbf{x}(k, N)$ the kth frame of the Nth network retraining. Index k is reset each time a new retraining phase takes place. Therefore, $\mathbf{x}(0, N)$ corresponds to the image on which the Nth retraining was accomplished. Fig. 2 indicates a scenario with four retraining phases of a video sequence composed of 300 frames; at frame 27 where shot 2 begins, at frame 138 where the start of shot 3 is detected, at frame 211 inside shot 3, and at frame 251 where shot 4 begins. The corresponding values of indexes k and N are also depicted. As is observed, for the third shot two alarm signals are activated, possibly denoting a shot with substantial changes.

In Sections VA and VB, the scene change detection module and the operational environment change module are described.

A. Scene Change Detection Module

The shot cut detection module comprises one of the two parts of the decision mechanism. Scene change detection is a very interesting problem in the field of video analysis and several attempts have been reported in the literature, which deal with the detection of cut, fading or dissolve changes either in the compressed or uncompressed domain [30], [31]. In our approach, the algorithm proposed in [30] has been adopted for detecting shots due to its efficiency and low computational complexity. This is due to the fact that it is based on the dc coefficients of the direct cosine transform (DCT) transform of each frame. These coefficients are directly available in the case of intracoded frames (I frames) of MPEG compressed video sequences, while for the intercoded ones (P and B frames), they can be estimated by the motion compensated error with minimal decoding effort.

B. Operational Environment Change Module

The operational environment change module (OECM) comprises the second part of the decision mechanism, responsible for detecting changes within shots. Toward this direction and since the correct segmentation mask, or equivalently, the desired outputs are not known in the unsupervised segmentation, it is necessary to provide an estimate of the segmentation error. This is achieved by the OECM through comparison of the initial VO-masks, as provided by the retraining set extraction module and the following masks, provided by the neural-network classifier.

In particular, let us assume that, after retraining, the neural network is applied to the following frames of the sequence. In this case, it is expected that the classifier will provide a segmentation mask of good quality for each VO, as long as no rapid changes in the operational environment occur. Consequently, for each VO, the difference between its initial segmentation mask, provided at the retraining time instance and each mask of the following frames should change small.

Let us denote by $V_i(k, N)$ a mask that refers to the ith VO of the frame $\mathbf{x}(k, N)$. Let us also denote by $\hat{V}_i(N)$ the approximation of the ith object, as provided by the N retraining set at

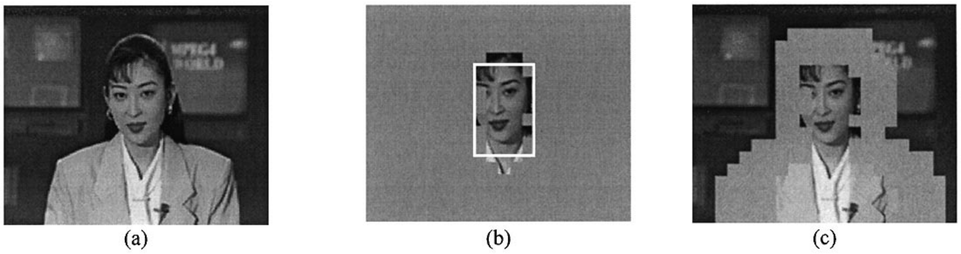


Fig. 3. Initial VO estimation for the Akiyo sequence. (a) Original frame. (b) Output of the face detection module. (c) Retraining set constructed for the foreground and the background object along with the region of uncertainty.

the Nth retraining. For each VO $V_i(k, N)$, several descriptors are extracted to describe the object properties, such as the color and texture histogram. The same descriptors are also extracted for the initial estimate of the ith Vo $\hat{V}_i(N)$. Then, the difference

$$e\left(V_i, 0, N\right) = \left\| \theta\left(V_i(0, N)\right) - \theta\left(\hat{V}_i(N)\right)_2 \right\|_2 \tag{19}$$

expresses the approximation error of the retraining set construction method for the ith VO at the frame, where retraining has been activated. In (19), $\boldsymbol{\theta}(\cdot)$ is the feature vector of the respective VO, while $\|\cdot\|_2$ indicates the L_2 norm.

Let us now denote by $e(V_i, k, N)$ the approximation error obtained at the kth video frame of the sequence after the Nth retraining. This error is provided by the following equation:

$$e\left(V_i, k, N\right) = \left\| \theta\left(V_i(k, N)\right) - \theta\left(\hat{V}_i(N)\right)_2 \right\|_2. \tag{20}$$

It is anticipated that the level of improvement provided by $e(V_i, k, N)$ will be close to that of $e(V_i, 0, N)$ as long as the classification results are good. This will occur when input images are similar, or belong to the same scene with the ones used during the retraining phase. An error $e(V_i, k, N)$, which is quite different from $e(V_i, 0, N)$, is generally due to a change of the environment. Thus, the quantity

$$a\left(V_i, k, N\right) = \frac{\left|e\left(V_i, k, N\right) - e\left(V_i, 0, N\right)\right|}{e\left(V_i, 0, N\right)} \tag{21}$$

can be used for detecting the change of the environment or equivalently the time instances where retraining should occur. Thus

$$\text{if } a(k, N) < T \text{ no retraining is needed} \tag{22}$$

where T is a threshold which expresses the maximum tolerance, beyond which retraining is required for improving network performance. In case of retraining, index k is reset to zero while index N is incremented by one. Using this criterion, an alarm signal is generated to preserve deterioration of the neural classifier's performance in several interesting cases. Such cases include disappearance of one or more of the VOs from a shot, or appearance of new VOs.

VI. Experimental Results

In this section, we evaluate the performance of the proposed unsupervised VO segmentation and tracking scheme in both of the examined scenarios. In particular, for the first scenario, three different video conferencing sequences were investigated. The first two are the Akiyo and Silent sequences adopted in the MPEG-4 and have been selected to indicate the ability of the algorithm to extract humans even in a complex background. The third presents a person who is bending and has been selected for investigating a complicated object motion situation. For the second scenario, the "Eye to Eye" stereoscopic sequence has been used. This sequence was produced in the framework of the ACTS MIRAGE project in collaboration with AEA Technology and ITC and is of total duration of 25 minutes.

In the performed experiments, each image pixel is classified to a VO according to a feature vector extracted from a block of size 8×8 pixels, centered around the respective pixel to be classified. Such a selection reduces possible noise in classification. As a result, VO segmentation is performed by considering overlapping blocks of 8×8 pixels. More specifically, the dc coefficient followed by the first eight zig-zag scanned ac coefficients of the DCT for each color component (i.e., $9 \times 3 = 27$ elements) are used to form the feature vector of the respective block. The neural-network architecture is selected to consist of $q = 15$ hidden neurons and three outputs, each of which is associated with a specific VO, since for simplicity, in the following experiments we assume that the maximum number of VOs is three, i.e., two foreground objects at maximum and the background. The case of more VOs can be considered in a similar manner. Therefore, the network has 450 network weights. Initially, a set of approximately 1000 image blocks is used to train the neural network. In the following, the two examined scenarios are separately discussed.

A. Case of Videophone/Videoconference Applications

In the first scenario, the three video sequences are put one after the other to construct a video conferencing stream of different shot content. Threshold T used in the decision mechanism [see (22)] was selected to be 30% so that alarm signals are generated only when substantial variations occur. Since in videoconferencing, the operational environments usually do not present abrupt changes, the retraining module was activated only at shot variations. Therefore, three retraining phases are detected.

At each retraining phase, an approximate estimation of the human entity is provided by applying the human face and body detection scheme, as described in Section III. Figs. 3 and 4 illustrate the initial VO estimation for the first two retraining phases, accomplished at the first frame of Akiyo and Silent sequence. These frames are presented in Figs. 3 and 4(a), while Figs. 3 and 4(b) indicate the performance of the human face detection module for the respective images. For clarity of presentation, when a block is classified to foreground, it is included as it is in the figures, while blocks classified to the background, are depicted with gray color.

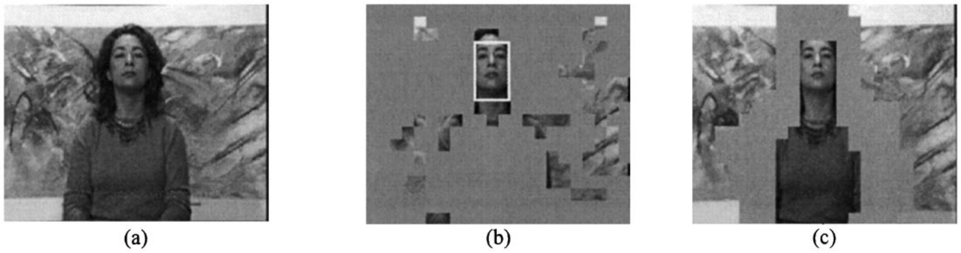


Fig. 4. Initial VO estimation for the Silent sequence. (a) Original frame. (b) Output of the face detection module. (c) Retraining set constructed for the foreground and the background object along with the region of uncertainty.

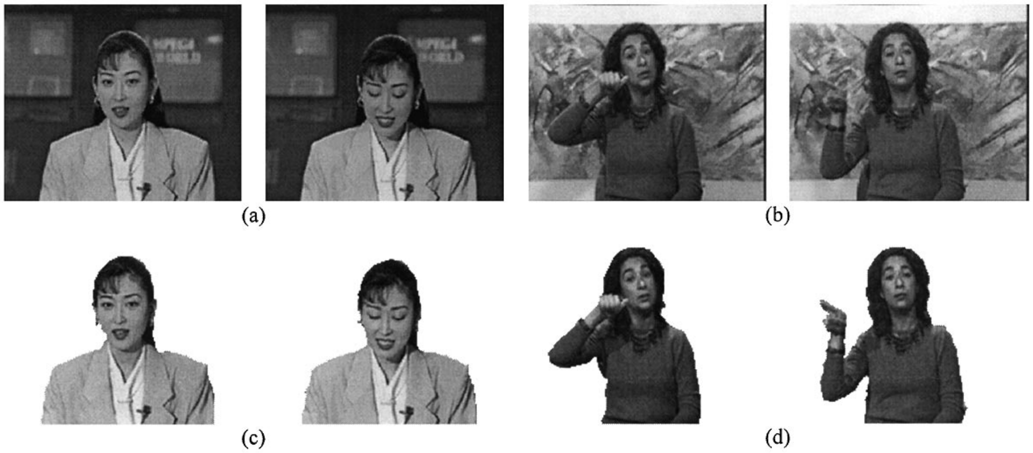


Fig. 5. Tracking performance. (a) Original frames 20 and 102 of the Akiyo. (b) Original frames 50 and 120 of the Silent. (c) Tracking performance Akiyo. (d) Tracking performance of Silent.

As can be observed, in Silent, additional blocks are also selected as candidate face blocks, due to the fact that their chrominance characteristics are close to those of face region. To eliminate false blocks, a template-matching algorithm is applied, as described in Section III-A, which exploits shape properties of human faces. For template matching, a rectangle is used of certain aspect ratio, which models the human face area. In our implementation, the aspect ratio lies in the interval [1.2 1.7]. The extracted human face, after the template matching procedure, is depicted in Figs. 3 and 4(b) as a white rectangle. As is observed, nonface regions are discarded (false alarms), as they do not satisfy the shape constraints, described in Section III-A.

In the following, an approximation of the human body location is performed to construct the retraining set used to update the weights of the adaptable neural-network architecture. The human body estimation is depicted in Figs. 3 and 4(c) for the same frames of the examined sequences. In these figures, we have also presented the retraining sets. In particular, blocks of gray color indicate regions of uncertainty, meaning that these blocks are not included in the retraining set. As can be observed, the region of uncertainty contains several blocks, located at the boundaries of the face and body areas and, thus, protects the network from ambiguous regions. Instead, the remaining blocks are included in the retraining set as foreground/background data. For each selected block, a feature vector is constructed based on the DCT coefficients of the block, as mentioned above, which are then used to retrain the neural network.

Since, however, there is a large number of blocks with similar content, a principal component analysis (PCA) is applied to reduce their number. More specifically the number of selected retraining blocks was reduced to ten in case of Akiyo and nine in case of Silent, after the application of the PCA analysis.

Fig. 5 depicts the tracking performance of the proposed scheme for two frames of the Akiyo (frames 20 and 102) and the Silent (frames 50 and 120) sequence. As is observed, the foreground object is extracted with high accuracy in all cases, despite the complexity of the background and the object motion. It can also be observed that the proposed method provides very satisfactory segmentation results for all different poses of the human objects as shown in Fig. 5(d) in which the Silent opens her arm.

Fig. 6 illustrates the performance of the proposed scheme to four characteristic frames of the third video conferencing sequence. In this sequence, a complicated situation is presented where a person is bending. As can be seen, the adaptable neural-network architecture accurately extracts the human object even in this complex case. This is due to the fact that the initial training set provides satisfactory information for identifying the human object, which is independent from the human location and orientation. Instead, deterioration of the network performance would be accomplished in case of a significant change of the shot content (e.g., change of the background characteristics). In such scenarios, however, a new retraining is activated to improve network performance and thus again, an accurate extraction of the human object is provided.

B. Exploitation of Depth Information

In the second scenario, the threshold T of the decision mechanism [see (22)], was selected to be 40%. As can be seen, the threshold is slightly greater than that of the first scenario, since we refer to generic sequences with more complicated content.

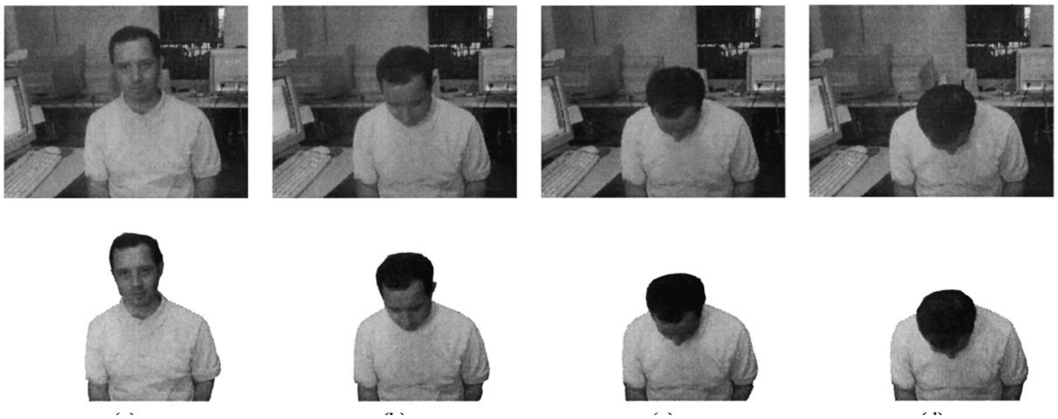
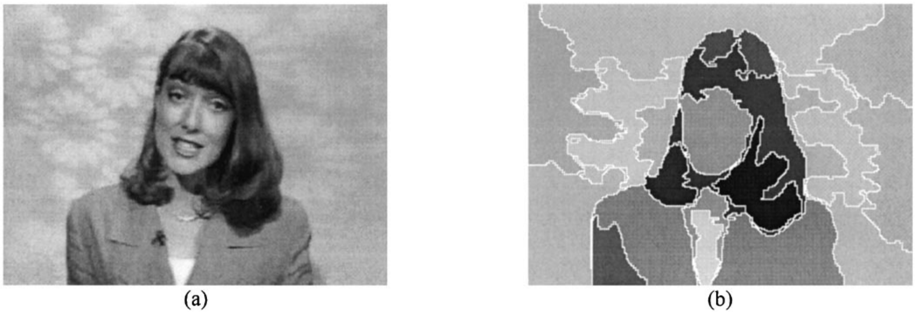


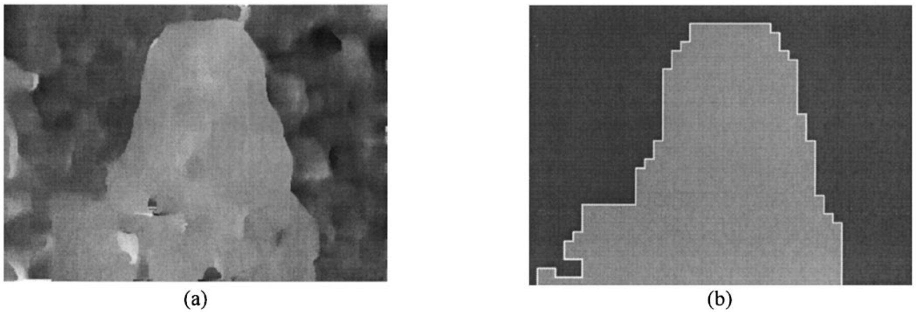
Fig. 6. Tracking performance on the “bending” sequence. (a) Four original characteristic frames. (b) Tracking performance.



(a)

(b)

Fig. 7. Color segmentation results for one frame of the “Eye to Eye” sequence. (a) Original left image. (b) Final color segmentation mask.



(a)

(b)

Fig. 8. Depth segmentation results for the image of Figure Fig. 7(a). (a) Occlusion compensated depth map. (b) Respective depth segments mask at the lowest resolution level of the M-RSST segmentation algorithm.

Let us assume that the decision mechanism activates a new retraining process at the stereo frame, depicted in Fig. 7(a). Then, this stereo frame is analyzed and a color and depth segmentation mask is constructed using the multiresolution recursive shortest spanning tree (M-RSST) segmentation algorithm described in [25] due to its efficiency and low computational complexity. The color segmentation is depicted in Fig. 7(b) where color segments accurately describe the boundaries of all different color regions. Considering now depth segmentation, an occlusion compensated depth map is first estimated as in [25], depicted in Fig. 8(a), while the respective segmentation mask is shown in Fig. 8(b). It should be mentioned that depth segmentation is performed only at the lowest resolution level of the M-RSST, since the depth map does not provide accurate VO boundaries, even if higher resolution levels are processed. This consideration justifies the block resolution of Fig. 8(b).

Then, segmentation fusion is incorporated to construct the retraining set. More specifically, color segments are projected onto the depth segments and then fused according to depth similarity. This is described in Fig. 9(a) for the image of Fig. 7(a). In particular, depth segmentation, shown with two different gray levels as in Fig. 8(b), is overlaid in Fig. 9(a) with the white contours of the color segments, as obtained from Fig. 7(b). After the fusion, accurate extraction of the foreground object is accomplished as illustrated in Fig. 9(b) and (c).

Based on color-depth segmentation fusion results, the retraining set S_c is constructed. Since there is a large number of similar retraining data, a PCA is used to reduce their number, as happens in the first scenario. More specifically, the number of selected retraining blocks was reduced to ten. Afterwards, weight adaptation of the neural classifier is accomplished, based on the algorithm described in Section II. Then, the

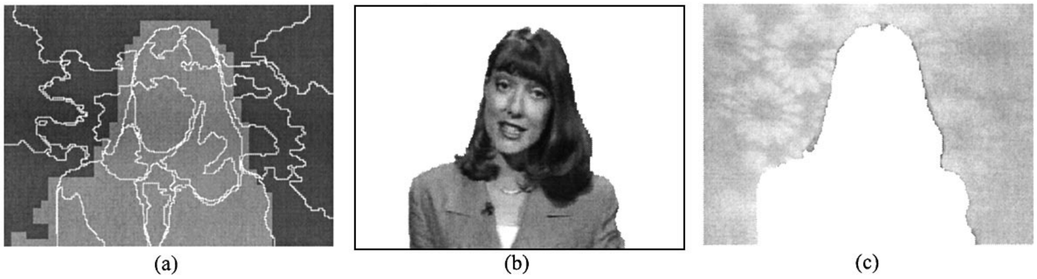


Fig. 9. Segmentation fusion results for the image of Fig. 7(a). (a) Depth segmentation overlaid with color segment contours (in white). (b) Foreground object. (c) Background object.

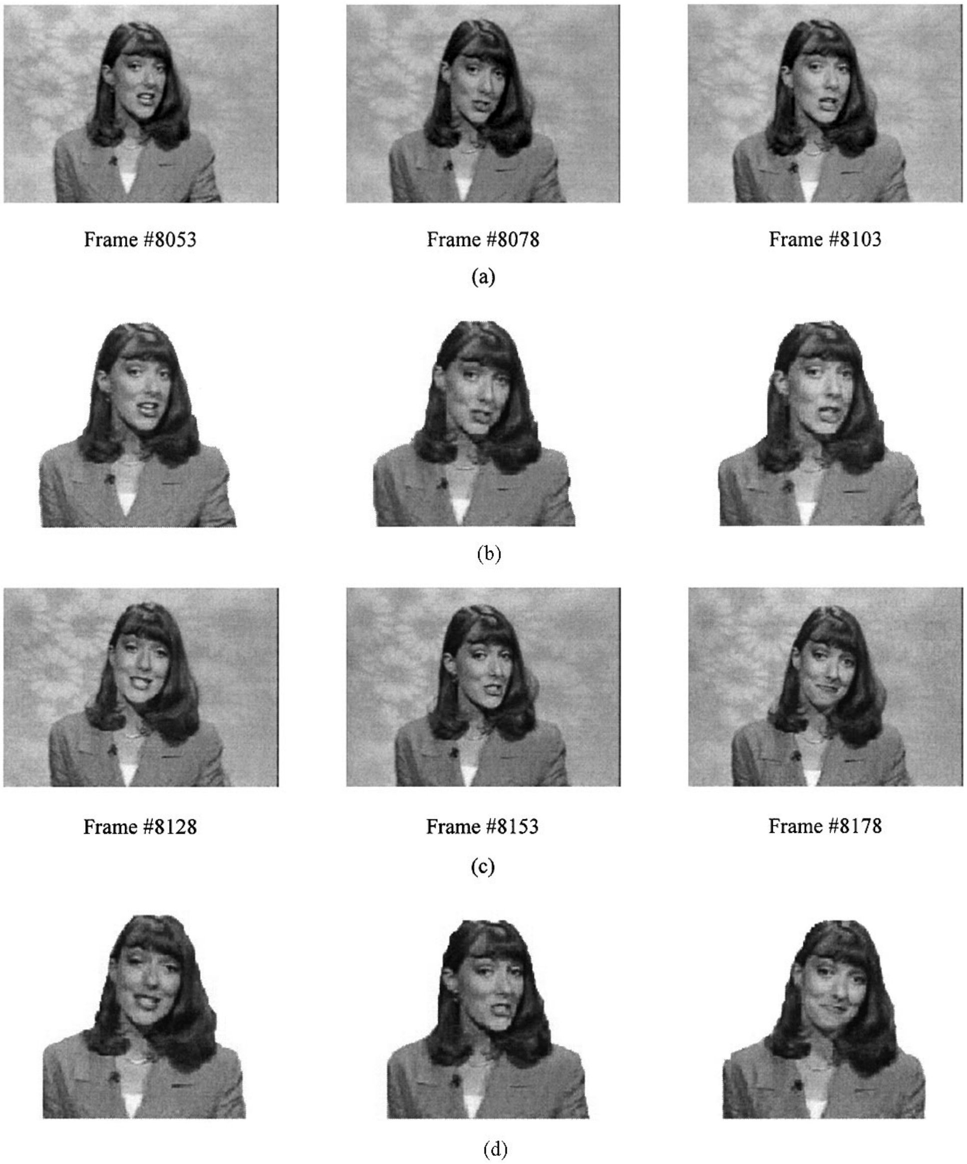


Fig. 10. Segmentation performance of the proposed adaptive neural-network architecture for several frames of the shot of Fig. 7 (first retraining phase). (a), (c) The original frames. (b), (d) The respective segmentation mask of the foreground object.

retrained network performs VO segmentation (the object tracking), until a new retraining process is activated.

In order to evaluate the segmentation efficiency of the network, we present in Fig. 10(b) and (d) the segmentation results of six different frames belonging to the same retraining interval with the frame of Fig. 7(a). We also depict the original left-channel stereo frames in Fig. 10(a) and (c). It should be mentioned that the VO (the speaker) is extracted in pixel resolution, since overlapping blocks are used.

The fifth retraining phase is activated for the stereo frame depicted in Fig. 11(a). The initial VO estimation is again performed by the color-depth segmentation fusion and presented in Fig. 11. Particularly, Fig. 11(b) depicts the color segmentation, while Fig. 11(c) and (d) illustrates the depth map and

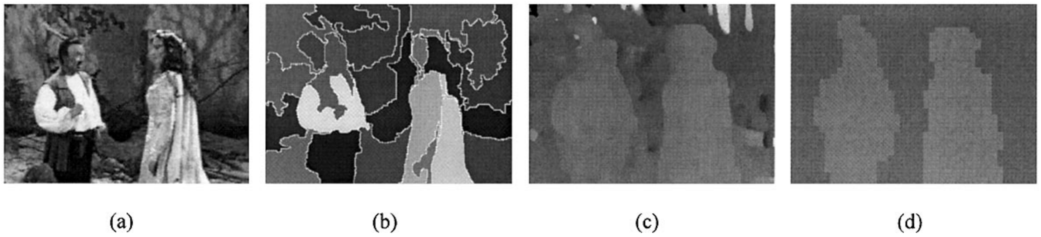


Fig. 11. Color-depth segmentation results for another frame of the "Eye to Eye" sequence. (a) Original left image channel. (b) Final color segmentation mask. (c) Depth map. (d) Respective depth segmentation.

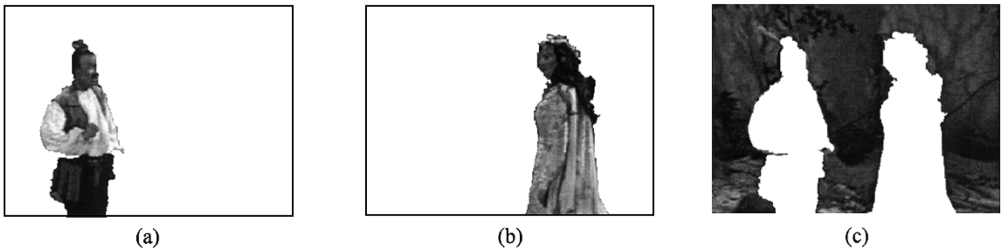


Fig. 12. Segmentation fusion results for the image of Fig. 11(a). (a) First foreground object (the "man"). (b) Second foreground object (the "woman"). (c) Background object.

the respective depth segmentation. Fig. 12 presents the initial extracted VOs, based on the color-depth segmentation fusion scheme, which are used, after applying a PCA analysis, to retrain the network. As can be seen, three VOs are extracted which correspond to the two persons (actors) and the background.

Fig. 13 presents the tracking results for four different characteristic frames of the fifth retraining phase. In this figure, the respective original left stereo channels are also depicted to provide an indication of the complexity of the visual content. As is observed, VO segmentation is remarkably performed by the retrained neural network. In particular, in this scene, the two actors are moving toward each other in a theatrical action. The adaptable neural network accurately tracks the two actors, even in cases that one is occluded from the other [see Fig. 13]. This is due to the fact that the initial retraining set provides sufficient information of the shot content. Therefore, as the background and the actors' characteristics remain the same, the adaptable neural-network architecture can efficiently extract the objects. Instead, using a motion-based tracking scheme as is described in the following, it is difficult to track the contours of the two objects in such complicated situations, since many errors appear mainly due to the occluded regions. In the proposed scheme, VO tracking is performed based on content characterization provided by the initial retraining set. Consequently, tracking is robust to complicated object motion or occlusion in contrast to conventional tracking algorithms.

C. Objective Evaluation

In the previous sections, the performance of the proposed scheme is evaluated on a subjective basis by depicting characteristic frames of complicated content. Instead, in this section, an objective criterion is introduced for measuring the quality of a segmentation/tracking algorithm. Objective evaluation criteria can be found in the MPEG-4 standard, within the core-experiment on automatic segmentation of moving objects [32]. Such quality measures provide a general framework for comparing different VO segmentation and tracking algorithms and ranks them according to their efficiency.

The objective evaluation assumes that *a priori* known (i.e., reference) segmentation mask exists and then, estimates the segmentation quality by comparing the shape of the reference mask with the shape of the one obtained by the algorithm to evaluate. In particular, the segmentation accuracy is measured through a mask error value. Two types of errors can be distinguished. The first corresponds to missing foreground points (the MF error), while the second to added background points (the AB error) [33]. As is observed, the MF error includes the points which although actually belong to the foreground object, they have been classified to the background. Therefore, the error MF is the number of pixels of set $R^c \cap S^c$, where R refers to the original (reference) mask and S to the segmented mask. The (^{c}) indicates the complement of set S. Similarly, the AB error is defined as the points which have been classified to the foreground object, while actually belong to the background. Thus, the AB error is expressed by the cardinality of set $R^c \cap S$. Taking into consideration both the MF and the AB errors, the total segmentation error is given as

$$E = \frac{\text{card}(R \cap S^c) + \text{card}(R^c \cap S)}{\text{card}(R)} \tag{23}$$

where $card(\cdot)$ refers to the cardinality (i.e., number of pixels) of a set. In order to homogenize the presentation with the SNR-like quality measures the error E is provided in a logarithmic scale as

$$Q = 10 \log \left(\frac{1}{1+E} \right). \tag{24}$$

1) Simulation and Comparisons: In this section, the segmentation error Q of (24) is used to objectively evaluate the performance of the proposed scheme in both the investigated

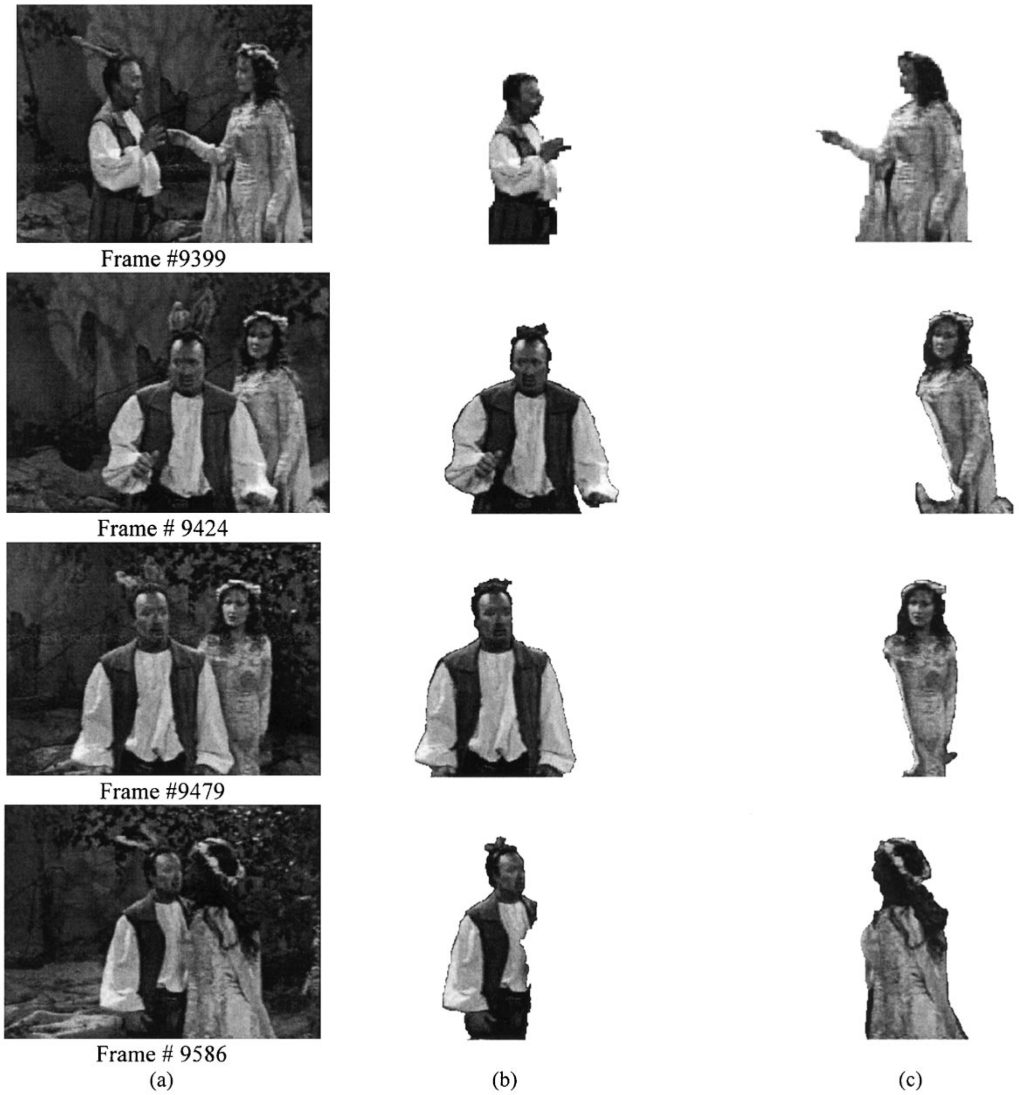


Fig. 13. Tracking performance at fifth retraining instance (Fig. 11). (a) Original frames. (b) Tracking of the first foreground object ("man"). (c) Tracking of the second foreground object ("woman").

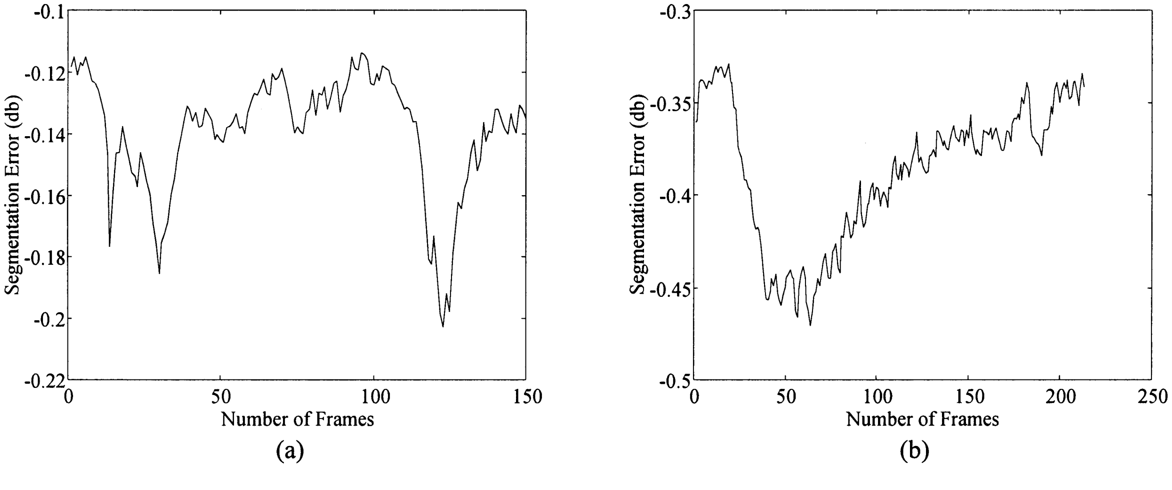


Fig. 14. Segmentation error Q, expressed in decibels, with respect to frame numbers. (a) Silent sequence (first scenario). (b) Fig. 11 (second scenario).

scenarios and to compare the adaptable neural-network architecture with other tracking/segmentation techniques. Fig. 14(a) presents the segmentation error, expressed in dB for the first 150 frames of the Silent sequence (first scenario). In this experiment, the initial VO has been unsupervisedly estimated by applying the algorithm described in Section III on the first frame

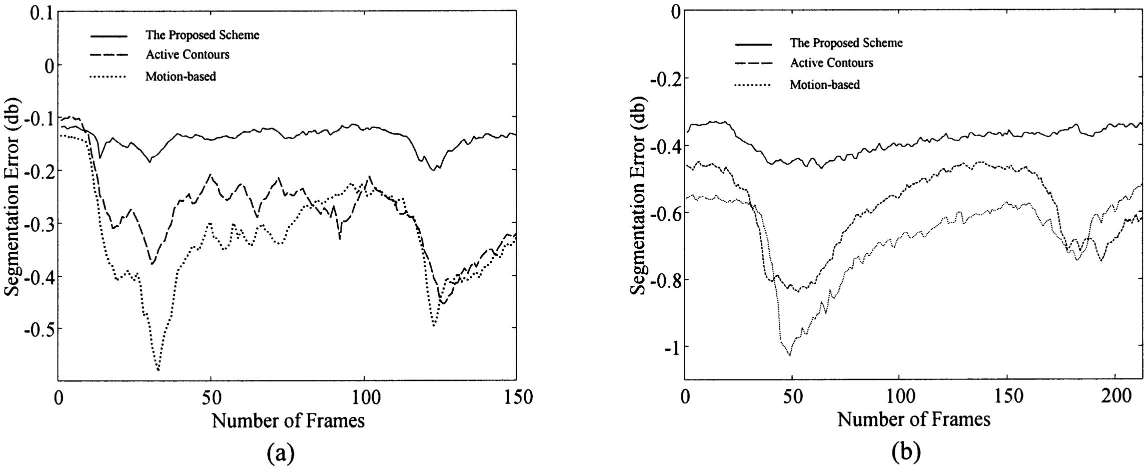


Fig. 15. Comparisons of the proposed scheme with an active contour and a motion-based method. (a) Silent sequence (first scenario). (b) Fig. 11 (second scenario).

of the Silent. This initial VO estimate is used to construct the retraining set S_c, which describes the current shot content and then the adaptable neural network performs object segmentation/tracking. As is observed, the tracking performance is very accurate despite, the fact that the sequence is characterized by complicated motion and a large number of frames are used. In particular, the worst performance is smaller than -0.22 dB, or about 5% error in the segmented mask. Furthermore, the highest segmentation error is noticed around 15–40 and around 120 frames since in this case, the Silent moves her hand up. Fig. 14(b) presents the algorithm performance for the first 213 frames of the shot of fifth retraining (see Fig. 11) of the second scenario. In this case, the evaluation has been performed for "Woman" object. As is observed, the proposed scheme accurately tracks the "woman," though in some frames she is occluded by the "man" (frames of worst performance).

Fig. 15 compares the proposed scheme with two other methods. The first uses active contours attracted by a gradient vector filed (GVF) [17], while the second deforms object boundaries by exploiting motion information like the ones presented in [34]. In particular, Fig. 15(a) shows the results for the Silent (first scenario), while Fig. 15(b) for the first 213 frames of respective shot of Fig. 11 (second scenario). In both cases, the proposed scheme outperforms the compared ones in segmentation and tracking accuracy. The compared techniques require an accurate initial VO segmentation, which cannot be given by the methods described in Section III and Section IV, which provide only an approximate (coarse) initial estimation of VOs. For this reason, in our implementation, the initial segmentation is manually extracted resulting in a semi-automatic scheme. This is also a drawback of these techniques compared to the proposed one where no user's interaction is required. Furthermore, they are sensitive to high motion changes and content variations. Instead, the proposed scheme captures the content characteristics (through the retraining set) and, thus, it is robust to motion fluctuations. In addition, the compared methods present an unstable behavior as the number of tracked frames increases and, therefore, new initializations are required. It should be also mentioned that the computational complexity of both compared techniques is much higher than the proposed one, since in our case VOs are tracked during the neural-network testing phase.

VII. Conclusion

New multimedia applications addressing content-based image retrieval, video summarization and content-based transmission, require an efficient representation of the image visual information. For this reason, Vo extraction has created a new challenging research direction and received great attention in the late years, both by the academic society and the industry. Several obstacles should be confronted till a generally applied scheme is produced, as VO segmentation in real-life video sequences is a difficult task, due to the fact that object characteristics frequently change through time.

In this paper, an unsupervised Vo segmentation/tracking algorithm is proposed using an adaptable neural-network architecture. The adaptive behavior of the network is very important in such dynamically changing environments, where object properties frequently vary through time. For the weight updating, a retraining set is constructed, which describes the content of the current environment. Then, network adaptation is performed so that 1) the network response, after the weight updating, satisfies the current environment and 2) the obtained network knowledge is minimally deteriorated.

The proposed unsupervised video segmentation/tracking scheme is investigated in two scenarios. The first deals with the extraction of human entities in videoconference sequences, while the second exploits depth information to segment generic semantically VOs. In the first scenario, a face and body detection module is incorporated, based on Gaussian distribution models and a binary template matching technique, while in the second case a color/depth segmentation fusion algorithm is presented for the initial Vo estimation. A decision mechanism is also incorporated in the proposed scheme to automatically detect the time instances that a new weight updating is required.

Experimental results on real life video sequences are presented and indicate the reliable and promising performance of the proposed scheme, even in cases of complex backgrounds, for objects with complicated nonrigid motion, or shots containing multiple VOs or occlusions. The performance is measured both subjectively and objectively using a segmentation error criterion. Furthermore, comparisons with other motion-oriented tracking schemes reveals the robustness of the proposed architecture.

Acknowledgment

The authors wish to thank C. Girdwood, Project Manager of the ITC (Winchester), for providing the 3-D video sequence "Eye to Eye." The authors also thank the anonymous reviewers and the Associate Editor for their constructive comments, which helped in improving the quality and presentation of the paper.

References

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

[2] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, "Very low bit-rate coding of image sequences using adaptive regions of interest," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 928–934, Dec. 1998.

[3] B. Furht, S. W. Smoliar, and H. Zhang, *Video and Image Processing in Multimedia Systems*, 2nd ed. Norwell, MA: Kluwer, 1996.

[4] N. Doulamis, A. Doulamis, Y. Avrithis, K. Ntalianis, and S. Kollias, "Efficient summarization of stereoscopic video sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, pp. 501–517, June 2000.

[5] Meyer and S. Beucher, "Morphological segmentation," *J. Visual Commun. Image Representation*, vol. 1, no. 1, pp. 21–46, Sept. 1990.

[6] M. Kunt, A. Ikonomopoulos, and M. Kocher, "Second generation image coding techniques," *Proc. IEEE*, vol. 73, pp. 549–574, Apr. 1985.

[7] O. J. Morris, M. J. Lee, and A. G. Constantinides, "Graph theory for image analysis: An approach based on the shortest spanning tree," *Proc. Inst. Elect. Eng.*, pt. F, vol. 133, no. 2, pp. 146–152, Apr. 1986.

[8] W. B. Thompson and T. G. Pong, "Detecting moving objects," *Int. J. Comput. Vision*, vol. 4, no. 1, pp. 39–57, Jan. 1990.

[9] J. Wang and E. Adelson, "Representing moving images with layers," *IEEE Trans. Image Processing*, vol. 3, pp. 625–638, Sept. 1994.

[10] G. Avid, "Determining three-dimensional motion and structure from optical flow generated by several moving objects," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-7, pp. 384–401, July 1985.

[11] T. Meier and K. Ngan, "Video segmentation for content-based coding," *IEEE Trans. Cicuits Syst. Video Technol.*, vol. 9, pp. 1190–1203, Dec. 1999.

[12] M. Kim, J. G. Choi, D. Kim, H. Lee, M. H. Lee, C. Ahn, and Y.-S. Ho, "A VOP generation tool: Automatic segmentation of moving objects in image sequences based on spatio-temporal information," *IEEE Trans. Cicuits Syst. Video Technol.*, vol. 9, pp. 1216–1226, Dec. 1999.

[13] C. Gu and M.-C. Lee, "Semiautomatic segmentation and tracking of semantic video objects," *IEEE Trans. Cicuits Syst. Video Technol.*, vol. 8, pp. 572–584, Sept. 1998.

[14] F. Bremond and M. Thonnat, "Tracking multiple nonrigid objects in video sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 585–591, Sept. 1998.

[15] I. K. Sethi and R. Jain, "Finding trajectories of feature points in a monocular image sequence," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, pp. 56–73, Jan. 1987.

[16] Y. S. Yao and R. Chellappa, "Tracking a dynamic set of feature points," *IEEE Trans. Image Processing*, vol. 4, pp. 1382–1395, Oct. 1995.

[17] C. Xu and J. L. Prince, "Snakes, shapes and gradient vector flow," *IEEE Trans. Image Processing*, vol. 7, pp. 359–369, Mar. 1998.

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

[19] T.-Y. Kwok and D.-Y. Yeung, "Constructive algorithms for structure learning in feedforward neural networks for regression problems," *IEEE Trans. Neural Networks*, vol. 8, pp. 630–645, May 1997.

[20] R. Reed, "Pruning algorithms—A survey," *IEEE Trans. Neural Networks*, vol. 4, pp. 740–747, Sept. 1993.

[21] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[22] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, "Learnability and the Vapnik-Chervornenkis dimesion," *J. Assoc. Computing Machinery*, vol. 36, no. 4, pp. 929–965, Oct. 1989.

[23] B. Cheng and D. M. Titterington, "Neural networks: A review from a statistical perspective," *Statist. Sci.*, vol. 9, pp. 2–54, 1994.

[24] A. Doulamis, N. Doulamis, and S. Kollias, "On line retrainable neural networks: Improving the performance of neural networks in image analysis problems," *IEEE Trans. Neural Networks*, vol. 11, Jan. 2000.

[25] A. Doulamis, N. Doulamis, K. Ntalianis, and S. Kollias, "Efficient unsupervised content-based segmentation in stereoscopic video sequences," *J. Artificial Intell. Tools*, vol. 9, no. 2, pp. 277–303, June 2000.

[26] D. J. Luenberger, *Linear and Non-Linear Programming*. Reading, MA: Addison-Wesley, 1984.

[27] D. Reisfeld, H. Wolfson, and Y. Yeshurum, "Detection of interest points using symmetry," in *Proc. Int. Conf. Coputer Vision*, Japan, Dec. 1990, pp. 62–65.

[28] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 696–710, July 1997.

[29] H. Wang and S.-F. Chang, "A highly efficient system for automatic face region detection in MPEG video sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 615–628, Aug. 1997.

[30] B. L. Yeo and B. Liu, "Rapid scene analysis on compressed videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, pp. 533–544, Dec. 1995.

[31] N. V. Patel and I. K. Sethi, "Video shot detection and characterization for video databases," *Pattern Recognition*, vol. 30, no. 4, pp. 583–592, Apr. 1997.

[32] M. Wollborn and R. Mech, "Procedure for Objective Evaluation of VOP Generation Algorithms,", Fribourg, Switzerland, Doc. ISO/IEC JTC1/SC29/WG11 MPEG97/2704, 1997.

[33] P. Villegas, X. Marichal, and A. Salcedo, "Objective evaluation of segmentation masks in video sequences," in *Proc. Workshop Image Analysis Multimedia Interactive Services*, Berlin, Germany, May–June 1999, pp. 85–88.

[34] M. Tekalp, *Digital Video Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

Anastasios Doulamis (S'96–M'00) received the Diploma (with the highest honor) and the Ph.D. degrees in electrical and computer engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 1995 and 2000, respectively.

From 1996 to 2000, he was with the Image, Video, and Multimedia Lab of the NTUA as a Research Assistant. From 2001 to 2002, he served his mandatory duty in the Greek army in the computer center department of the Hellenic Air Force, while in 2002, he join the telecommunication laboratory of the NTUA as senior researcher. He is author of more than 100 papers in the above areas, in leading international journals and conferences. His research interests include, non-linear analysis, neural networks, multimedia content description, and intelligent techniques for video processing.

Dr. Doulamis has received several awards and prizes during his studies, including the Best Greek Student in the field of engineering in national level in 1995, the Best Graduate Thesis Award in the area of electrical engineering with A. Doulamis in 1996, and several prizes from the National Technical University of Athens, the National Scholarship Foundation and the Technical Chamber of Greece. His Ph.D. work was supported by the Bodosakis Foundation Scholarship. In 1997, he received the NTUA Medal as Best Young Engineer. In 2000, he received the best Ph.D. dissertation award by the Thomaidion Foundation with N. Doulamis. In 2001, he served as Technical Program Chairman of the VLBV'01. He has also served on the program committee in several international conferences and workshops. He is Reviewer of IEEE journals and conferences as well as and other leading international journals.

Nikolaos Doulamis (S'96–M'00) received the Diploma degree (with the highest honor) and the Ph.D. degrees in electrical and computer engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 1995 and 2000, respectively.

He joined the Image, Video, and Multimedia Lab of NTUA in 1996 as a Research Assistant. From 2001 to 2002, he served his mandatory duty in the Greek army in the computer center department of the Hellenic Air Force. Since 2002, he has been a Senior Researcher in the communication lab of the NTUA. His research interest include video transmission, content-based image retrieval, summarization of video sequences and intelligent techniques for video processing.

Dr. Doulamis was awarded as the Best Greek Student in the field of engineering in national level by the Technical Chamber of Greece in 1995. In 1996, he was received the Best Graduate Thesis Award in the area of electrical engineering with A. Doulamis. During his studies, he has also received several prizes and awards from the National Technical University of Athens, the National Scholarship Foundation and the Technical Chamber of Greece. His Ph.D. work was supported by the Bodosakis Foundation Scholarship. In 1997, he was given the NTUA Medal as Best Young Engineer. In 2000, he was served as Chairman of technical program committee of the VLBV'01 workshop, while he has also served as program committee in several international conferences and workshops. In 2000, he was given the Thomaidion Foundation best journal paper award in conjunction with A. Doulamis. He is editor of the Who's Who bibliography. He is a Reviewer of IEEE journals and conferences as well as and other leading international journals.

Klimis Ntalianis (S'99) was born in Athens, Greece, in 1975. He received the Diploma degree and the Ph.D degree in electrical and computer engineering, both from the National Technical University of Athens (NTUA), Athens, Greece, in 1998.

He is the author of more than 45 scientific articles, while his research interests include 3-D image processing, video organization, multimedia cryptography and data hiding.

During the last four years, Dr. Ntalianis has received six prizes for his academic achievements. His Ph.D studies were supported from the National Scholarships Foundation and the Institute of Communications and Computers Systems of the NTUA. Dr. Ntalianis is a member of the Technical Chamber of Greece.

Stefanos Kollias (M'86) was born in Athens, Greece, in 1956. He received the Diploma degree in electrical engineering from the National Technical University of Athens (NTUA) in 1979, the M.Sc. degree in communication engineering from the University of Manchester Institute of Science and Technology, Manchester, U.K., in 1980, and the Ph.D. degree in signal processing from the Computer Science Division of NTUA in 1984.

Since 1986, he has served as Lecturer, Assistant Professor, and Associate Professor the Department of Electrical and Computer Engineering of NTUA. From 1987 to 1988, he was a Visiting Research Scientist in the Department of Electrical Engineering and the Center for Telecommunications Research of Columbia University, New York, on leave from NTUA. Since 1997, he has been a Professor with NTUA and Director of the Image, Video, and Multimedia Systems Lab. His research interests include image and video processing, analysis, coding, storage, retrieval, multimedia systems, computer graphics and virtual reality, artificial intelligence, neural networks, human computer interaction and medical imaging. Fifteen graduate students have completed their Doctorate under his supervision, other ten currently pursuing their Ph.D. degree. He has published more than 180 papers, 80 of which are in international journals. The last few years, he and his team have been leading or participating in more than 50 projects, both European and National.

Dr. Kollias received an honorary Diploma in the Annual Panhellenic Competition in Mathematics. In 1982, he received a COMSOC Scholarship from the IEEE Communication Society.