Multiresolution video object extraction fitted to scalable wavelet-based object coding

F.A. Tab, G. Naghdy and A. Mertins

Abstract: To enable content based functionalities in video processing algorithms, decomposition of scenes into semantic objects is necessary. A semi-automatic Markov random field based multiresolution algorithm is presented for video object extraction in a complex scene. In the first frame, spatial segmentation and user intervention determine objects of interest. The specified objects are subsequently tracked in successive frames and newly appeared objects/regions are also detected. The video object extraction algorithm includes discrete wavelet transform decomposition multiresolution Markov random field (MRF)-based spatial segmentation with emphasis on border smoothness at different resolutions, and an MRF-based backward region classification that determines the tracked objects in the scene. Finally, a motion constraint, embedded in the region classifier, determines the newly appeared objects/regions and completes the proposed algorithm towards an efficient video segmentation algorithm. The results are applicable for generic segmentation applications, however the proposed multiresolution video segmentation algorithm supports scalable object-based wavelet coding in particular. Moreover, compared to traditional object extraction algorithms, it produces smoother and more visually pleasing shape masks at different resolutions. The proposed effective multiresolution video object extraction method allows for larger motion, better noise tolerance and less computational complexity.

1 Introduction

The increasing popularity of multimedia applications calls for the development of image and video processing methods for effective distribution and representation of the visual information to provide new image/video services, such as interactivity, manipulation, editing, content-based access and scalability. To achieve these demands, image/video processing has moved away from block-based towards object-based techniques. Object oriented processing provides the great flexibility needed for new content-based services such as interactivity and manipulation. To this end, industrial standards which support object-based representation of audiovisual information were introduced by the Moving Pictures Expert Group (MPEG) [1]. MPEG-4 and MPEG-7 provide flexibility in manipulation, interactivity, editing, easier archiving and content-based access and retrieval from audiovisual databases [1, 2].

To enable the object-based image and video processing, semantic segmentation which decomposes the scene into meaningful objects is essential. The most challenging aspect of this process is the fact that low-level features do not lead to semantic objects directly, because a generic object may contain different grey-levels, colours, textures, motions and so on. The gap between meaningful objects and low-level features makes automatic and comprehensive semantic segmentation a very difficult task. Although a great deal of research in segmentation has been carried out, no dominant solution for this task has emerged. The proposed methods, by and large, remain *ad hoc* with little underlying theoretical foundation. Furthermore, segmentation is inherently an ill-posed problem [3]. This means that there is no unique solution to solve the multi-faceted segmentation problem. There are many different segmentation algorithms designed for specific problems with some simplified assumptions. This makes segmentation algorithms application dependent. On the other hand, segmentation is a first stage of processing for many image/video processing applications such as pattern recognition, image analysis and understanding, computer vision, image and video databases with content-based access and object-based coding. In particular, the new advances in networking and digital processing offer the potential for an explosion in multimedia applications over networks which require enabling object-based processing. In conclusion, there is a wide area of segmentation applications. It is a very important and formidable task with high demands and requires a great deal of intensive research.

Although a large number of automatic or semi-automatic video object segmentation methods have been proposed [4–8], ideal segmentation is far from reality at this stage of technology and the scope of research in this topic is still very wide. The concern in this paper is two areas of research which the available segmentation algorithms have not been able to effectively resolve. Underpinning these two areas is the concept of (spatial) scalability, where the 'object-of-interest' is extracted at different resolutions of pyramid decompositions and visual quality is a constraint. In this paper, considering the importance of coding for information distribution over heterogeneous networks, special attention is given to the application of the proposed

doi:10.1049/iet-ipr:20045155

Paper first received 28th August 2004 and in final revised form 30th August 2006

F.A. Tab is with the Department of Electrical and Computer Engineering, University of Kurdistan, Sanandaj, Iran

G. Naghdy is with the School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW 2522, Australia

A. Mertins is with the Institute for Signal Processing, University of Lübeck, D-23538 Lübeck, Germany

E-mail: fardin.tab@gmail.com

segmentation algorithms with scalable wavelet-based object coding algorithms [9, 10], although the results are useful for generic segmentation applications such as pattern recognition, image understanding and computer vision.

In a network environment such as in the Internet, it is desirable that a large number of users with different processing capabilities and network access bandwidth could access and transfer data easily. A new challenge with such a heterogenous environment is to design a coding system that produces a single bitstream for a given source signal which is capable of optimally servicing each end user according to individual bandwidth and computing capabilities. To overcome this challenge, some sort of scalability needs to be provided by the encoder. In object/frame-based scalable coding, the bitstreams for low-end users are embedded as a subset of the codestreams for high-end applications. As a result, a single bitstream can be applied to different users by selectively transmitting and decoding the related parts of the bitstream [11, 12]. Some of the desirable scalable functionalities are signal to noise ratio (SNR) scalability, spatial scalability and temporal scalability [12]. In object-based spatial scalability, the shapes and their texture information are coded and decoded on the basis of a specific resolution. In this case, the resolution is determined in correspondence with the end user's capabilities such as bandwidth and display resolution. Therefore considering the spatial scalability of scalable object-based encoder/decoder system, it is necessary to extract objects' shape at different resolutions.

The existing video segmentation algorithms in the literature extract the shape at the highest resolution [4–8]. Therefore a regularly used option to produce shape at different resolutions is to extract objects at the highest resolution followed by downsampling. However, this single-resolution procedure fails to deal with the requirement of multiresolution scalable segmentation and extraction processes and loses the properties and advantages of multiresolution processing, such as less computational complexity, better capturing of the image structure and less noise sensitivity. Moreover, this method cannot assure to produce the best shapes for lower-resolutions for all shapes, and it can produce shapes that are visually less pleasing. Downsampling could result in deformation and distortion at low resolution, which can damage the semantics of the objects such as creation of holes in the object area [13].

In other words, a visually pleasing object at higher-resolution does not necessarily ensure similar quality at lower-resolutions. For example in Fig. 1, downsampling of two digital circles are compared. It can be seen that better approximation of a digital circle at high resolution can result in worse downsampled shape.

In assessing the performance of the segmentation processes, traditionally, the main emphasis is placed on the statistical accuracy, while qualities such as well-defined borders or visual merit of the extracted objects are not considered. Visual quality of the segmented objects, however, has great influence on the viewers. Therefore as well as the statistical criteria, visual effect and quality criteria should be incorporated into the segmentation algorithms. This paper presents a semi-automatic object extraction algorithm which produces enhanced and visually pleasing objects at different resolutions. To obtain more visually pleasing shapes, the region/object smoothness has been considered as a criterion in the segmentation process. Considering the multiresolution applications such as spatial scalable wavelet-based object coding algorithms, the visually pleasing criterion is extended to multiresolution object extraction and analysis.

The proposed algorithm includes a semi-automatic object extraction algorithm which is based on spatial segmentation and MRF-based backward region classification. The proposed spatial segmentation fits multiresolution Markov random field (MMRF) segmentation to scalable object-based wavelet coding [14]. The image at different resolutions is segmented with spatial scalability as a constraint. To extract enhanced shapes, border smoothness, as a criterion of shape analysis [15], is also included in the objective function of spatial segmentation. For optimisation of MMRF modelling, the iterated condition mode (ICM) algorithm [16] matched to the scalable multiresolution segmentation is used. The 'object-of-interest' is determined by user intervention at the first frame and is tracked by a backward MRF-based region classifier which determines the foreground regions in the subsequent frames. Motion constraint which determines the newly appeared objects/regions completes the proposed algorithm towards an efficient semi-automatic, multiresolution and semantic video segmentation algorithm.

2 Object-based wavelet decomposition

Because of the attractive features of wavelet-based coding schemes such as potential to support SNR, spatial and temporal scalability, spatial-frequency analysis support, high energy compaction in low frequency coefficients and consistency with the human visual system, wavelet-based image/video coding schemes have become increasingly important and have gained widespread acceptance. An example is the JPEG 2000 still image compression standard [17]. Finally, depending on the shape of filters used for the wavelet decomposition during the encoding procedure, there is an exact downsampling relationship between the

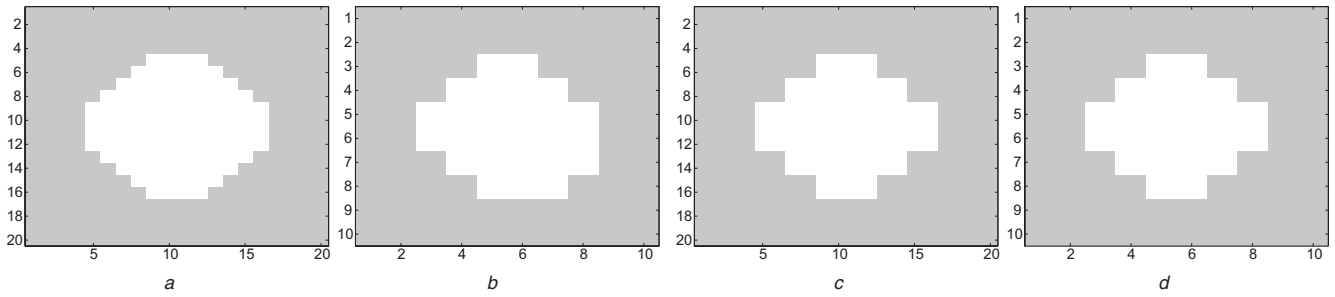

Fig. 1 *Circles in different resolutions*

a Closer approximation of a digital circle at high resolution
b Downsampling to low resolution
c Worse approximation of a digital circle at high resolution
d Downsampling of *c* to low resolution

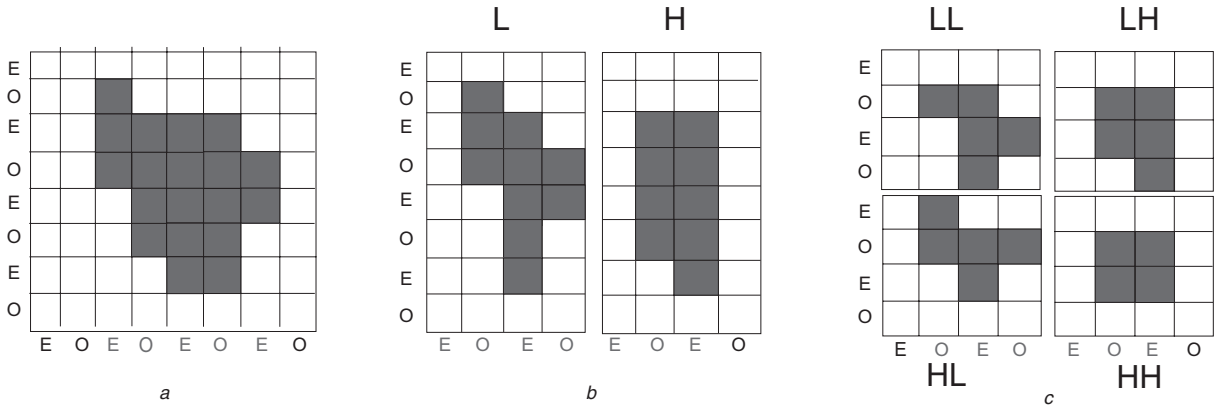

Fig. 2 *Decomposition of a non-rectangular object with odd-length filters*

a Object shown in dark grey
b Decomposed object after horizontal filtering
c Decomposed object after vertical filtering
'E' and 'O' indicate the position (even or odd) of a pixel in the horizontal and vertical dimensions

higher- and lower-resolution shapes [18–20]. In this paper, an odd-length filter (e.g. 9/7) is used, where all shape pixels with even indices (suppose indices start from zero or an even number) are downsampled for the lowpass band [18]. Fig. 2 illustrates the wavelet decomposition of an arbitrarily shaped object when using an odd-length filter. The final four-band decomposition is depicted in Fig. 2*c*. Considering the self-similarity of the wavelet transform, it is straightforward to suppose that the pixels of a shape with even index have the same segmentation classification as the corresponding pixels on its lower-resolution.

The wavelet self-similarity and down sampling relationship extends to all lowpass subband shapes of different resolutions. Therefore the discussed relationship between corresponding pixels is extended to shapes at different resolutions. Pixels with indices that are multiples of 2^n are downsampled to n lower-resolutions of the pyramid. These pixels are corresponding pixels at different levels. Therefore the number of downsampling or equivalently the number of corresponding pixels at different resolutions depends on the pixel indices. Moreover, any pixel at low resolutions has a corresponding pixel at the higher resolutions.

As mentioned earlier, the relationship between corresponding object pixels at different resolutions should be maintained and considered as a scalability constraint in the mask producing algorithms. Therefore considering scalability and wavelet self-similarity, a pixel and its corresponding pixels at all other levels have the same segmentation label. They only change together during segmentation.

3 MRF-based image segmentation algorithm

The main challenge in multiresolution image segmentation for scalable wavelet-based object coding is to keep the same relation between the extracted objects/regions at different resolutions as it exists between the decomposed objects at different resolutions in a shape-adaptive wavelet transform. The other constraint is border smoothness particularly in lower-resolutions. Different smoothness coefficients defined at different resolutions give some degree of freedom to put more emphasis on the low-resolution smoothness. To meet these challenges, Markov random field modelling is selected as it includes low-level processing at pixel level and has enough flexibility in defining objective functions tailored to the problem at hand [21]. In the following, first the principle of MRF-based single-resolution image segmentation is explained and then it is extended to the multiresolution scalable mode. Subsequently, a smoothness term is added to the objective criterion, and then the maximum a posteriori (MAP) estimation is presented.

3.1 Single-resolution image segmentation

In a regular single-level MRF-based image segmentation, the problem is formulated using a criterion such as the MAP criterion. If the desired segmentation is denoted by X and the observed intensity function is given by Y, according to the Bayes rule, the a posteriori probability can be written as

$$P(X|Y) \propto P(Y|X)P(X) \tag{1}$$

where $P(X|Y)$ represents the conditional probability of the segmentation label, given the observation, that is intensity value Y. Label field X is normally modelled by a MRF. Spatial continuity is easily incorporated into the segmentation, because it is inherent to MRFs [22]. Using a four- or eight-neighbourhood system considering only pairwise cliques, $P(X)$ is then a Gibbs distribution [23] and is defined by its energy function $U(X)$

$$P(X) = \frac{1}{Z}\exp\left(-\frac{1}{T}U(X)\right), \quad U(X) = \sum_{c\in C} V_c(X) \tag{2}$$

where Z and T are normalising constants and usually do not have to be evaluated. C is the set of all cliques, and V_c is the individual clique potential function. A clique is a set of neighbouring pixels. A clique function depends only on the pixels that belong to the clique. In single-level segmentation, usually one- or two-pixel cliques are used as shown in Fig. 3*a*, and for one-pixel cliques is assumed that the one pixel clique potentials are zero, which means that all region types are alike [23]. Spatial connectivity of the segmentation is imposed by assigning the following clique function:

$$V_c(X) = \begin{cases} -\beta & \text{if } X(i,j) = X(k,l) \text{ and } (i,j),(k,l) \in C \\ +\beta & \text{if } X(i,j) \neq X(k,l) \text{ and } (i,j),(k,l) \in C \end{cases} \tag{3}$$

where β is a positive number, and s and r are a pair of neighbouring pixels. Note that a low potential or energy

corresponds to a higher probability for pixel pairs to have identical labels. This encourages spatially connected regions.

To derive the conditional probability density $P(Y|X)$, the image is modelled as a collection of regions (a region is a set of connected pixels with the same label) with uniform or slowly varying grey-level. More precisely, the intensity of region m is modelled as a constant signal μ_m plus additive, zero mean white Gaussian noise with variance σ^2. The value of μ_m is computed by averaging the grey-level of all pixels belonging to region m in the current estimation of segmentation filed. In a more sophisticated model, the mean value $\mu_{X(s)}(s)$ is a slowly varying function of pixel s. The value of $\mu_{X(s)}(s)$ is computed by averaging the grey-level of all pixels at neighbouring of pixel s which also belong to the same region m in the current estimation of segmentation field. Therefore at each pixel s of region m, the image grey-level is characterised by a $\mu_{X(s)}(s)$ plus additive, zero mean white Gaussian [23, 24].

$$P(Y|X) \propto \exp\left(-\sum_s \frac{1}{2\sigma^2}(Y(s)-\mu_{X(s)}(s))^2\right) \quad (4)$$

Considering equations (1), (2) and (4), the probability density becomes

$$P(X|Y) \propto \exp\left\{-\left(\sum_s \frac{1}{2\sigma^2}(Y(s)-\mu_{X(s)}(s))^2 + \frac{1}{T}\sum_c V_c(X)\right)\right\} \quad (5)$$

Considering the MAP criterion, probability $P(X|Y)$ should be maximised which is equivalent to maximising the argument of the exponential function in equation (5) or minimising it's negative value. The argument of exponential function in (5) consists of two terms. Minimising the first term encourages the intensity function to be close to the estimated region's average. The second term encourages the adjacent pixels to have the segmentation label. Emphasis on any of these two terms can be adjusted by the value of any of three parameters σ, T and β. Therefore to simplify the expression, the parameters $2\sigma^2$ and T are set to one, and the segmentation result is controlled by the value of β in the V_c function. This results in the following cost or objective function which has to be minimised with respect to $X(s)$

$$E(X) = \sum_s \left\{\left(Y(s)-\mu_{X(s)}(s)\right)^2 + \sum_{r\in\partial s} V_c(s,r)\right\} \quad (6)$$

where ∂s denotes the set of neighbouring pixels of s.

To obtain the final segmentation, this objective function is minimised by one of the several MRF objective minimisation methods [21].

3.2 Multiresolution scalable image segmentation

In this section, the objective function of the single-resolution image segmentation algorithm is extended to a multiresolution scalable mode. To tailor the single-resolution objective function in (6) to our application, the wavelet transform is applied to the original image and a pyramid of decomposed images at various resolutions is created. Let Y denote the pyramid of grey-level pixels. The segmentation of the image into regions at different resolutions will be denoted by X. To change the segmentation label of a pixel, as explained in Section 2, the pixel and all its corresponding pixels at all other levels have to be analysed together. As a result, an analysis of a set of pixels in a multidimensional space, instead of a single-resolution analysis, needs to be used. Instead of speaking of a set of pixels, in the multidimensional space, the term 'vector' or 'array' is used for convenience (direction is not important). An array includes corresponding pixels at different resolutions of the pyramid. A symbol $\{\boldsymbol{s}\}$ shows an array which includes pixel s and its corresponding pixels at different resolutions. The length of an array is equal to the number of corresponding pixels at different resolutions, which depends on the index of pixels, and it can be 1, 2 or more. Therefore clique definition is extended to multidimensional or multiresolution space. The extended cliques act on two arrays instead of two pixels. Fig. 3*a* shows regular one- and two-pixel clique sets. In Fig. 3*b*, the extension of one of these cliques to the array mode can be seen.

The extension of clique functions is achieved through the following steps: (3) is used for cliques with length two at a resolution where pixels s and r are two neighbouring pixels on the same resolution level. Equation (7) is defined for multiple levels

$$\begin{aligned} V_c(\{\boldsymbol{s}\},\{\boldsymbol{r}\}) &= \left(\frac{1}{N}\right)\sum_{k=M}^{M+N-1} V_c(s_k,r_k) \\ &= \left(\frac{1}{N}\right)\sum_{k=M}^{M+N-1}(-1)^{L_k}\beta \\ L_k &= \begin{cases} 1 & \text{if } X(s_k)=X(r_k) \\ 0 & \text{if } X(s_k)\neq X(r_k) \end{cases} \quad s_k\in\{\boldsymbol{s}\},\ \ r_k\in\{\boldsymbol{r}\} \end{aligned} \quad (7)$$

where $\{\boldsymbol{s}\}$ and $\{\boldsymbol{r}\}$ are two neighbouring arrays which include two neighbouring pixels s and r. The neighbouring pixels of

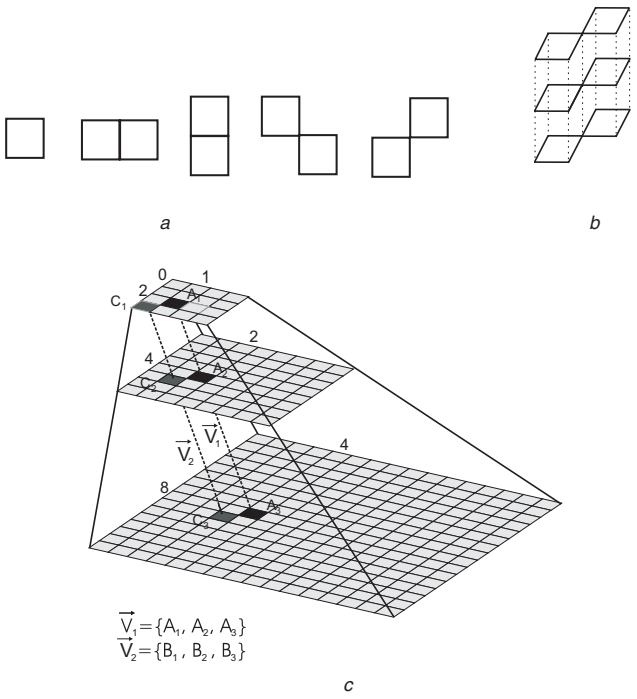

Fig. 3 *Single- and multiresolution image segmentation*

a Normal one- and two-pixel clique sets
b Clique of two vectors with the vectors' dimension equal to three
c Clique and its two neighbouring vectors $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ with dimension 3 are shown on the pyramid. Dashed lines connect the corresponding pixels of vectors

the two arrays $\{s\}$ and $\{r\}$ at level k are denoted s_k and r_k. The lowest resolution pixel on vector $\{s\}$ is denoted by M and the vector's dimension is denoted N. At each resolution such as level k, the single resolution clique function compares two neighboring pixels s_k and r_k and the value $+\beta$ or $-\beta$ is achieved. Therefore (7) is the average of single resolution clique function at different resolutions. A positive value is assigned to parameter β, so that two neighbouring pixels at the same level are more likely to belong to the same class than to different classes [23]. It is notable that (7) extends the clique definition to multiresolution mode. Therefore the objective function can be written as follows

$$E(X) = \sum_{\{S\}} \left\{ \| Y(\{s\}) - \mu_{X(\{s\})}(\{s\}) \|^2 + \sum_{\{r\} \in \partial\{s\}} V_c(\{s\}, \{r\}) \right\} \tag{8}$$

The first summation is over the set of different arrays on the pyramid, whereas the inner summation is over the set of all neighbouring arrays of vector $\{s\}$, denoted by the $\partial\{s\}$. The two arrays $\{s\}$ and $\{r\}$ are neighbours if pixels of $\{s\}$ and $\{r\}$ located at the same resolution are also neighbours. The grey-levels of pixels in set $\{s\}$ form an array $Y(\{s\})$. Similarly, $\mu(\{s\})$ and $X(\{s\})$ are mean and segmentation label arrays, respectively. The explained approach used in this section to develop (8) is a generalisation from regular to scalable multiresolution image segmentation algorithm.

3.3 Scalable colour-image segmentation

The proposed multiresolution scalable algorithm can be extended to colour images. Similar to grey-level image segmentation, first the objective function of single-resolution segmentation is extracted and then it is extended to scalable mode. Let Y be the observed colour image with three channels shown by a three dimensional vector $Y = [Y_1, Y_2, Y_3]$ and the desired segmentation be denoted by X. By assuming the conditional independence of the channels given the segmentation field [25], we have $P(Y|X) = P(Y_1|X)P(Y_2|X)P(Y_3|X)$. Then according to the Bayes rule, the a posteriori probability density of the segmentation variables can be written as the conditional probability

$$P(X|Y) \propto P(Y|X)P(X) = P(Y_1|X)P(Y_2|X) \times P(Y_3|X)P(X) \tag{9}$$

If each of the probability functions $P(Y_i|X)$, $i = 1, 2, 3$ can be shown by an equation similar to (5), then (9) at single-resolution mode can be written as

$$P(X|Y_i) \propto \exp\left\{ -\sum_s \left(\sum_{i=1}^{3} \frac{1}{2\sigma_i^2} (Y_i(s) - \mu^i_{X(s)}(s))^2 + \frac{1}{T} \sum_{r \in \partial_s} V_c(s, r) \right) \right\} \tag{10}$$

where $i = 1, 2, 3$ corresponds to the different colour channels. Parameters $2\sigma_i^2$, $i = 1, 2, 3$ and T can be set to one, similar to grey-level segmentation. Then, according to the MAP criterion, the objective function can be written as follows

$$E(X) = \sum_s \left(\sum_{i=1}^{3} (Y_i(s) - \mu^i_{X(s)}(s))^2 + \frac{1}{T} \sum_{r \in \partial_s} V_c(s, r) \right) \tag{11}$$

When moving from single-resolution to multiresolution scalable mode, Y_i, $i = 1, 2, 3$ in $\boldsymbol{Y} = [Y_1, Y_2, Y_3]$ denotes the intensities of different colour channels in the pyramid's pixels. Similarly, the multidimensional processing should be used and the corresponding pixels at different resolutions are expressed by the symbol $\{\}$. The concept of clique functions is the same as for the scalable grey-level image segmentation mode. Therefore, the objective function in (11) is extended to scalable colour mode as follows

$$E(X) = \sum_{\{S\}} \left\{ \sum_{i=1}^{3} \| Y_i(\{s\}) - \mu^i_{X(\{s\})}(\{s\}) \|^2 + \sum_{\{r\} \in \partial\{s\}} V_c(\{s\}, \{r\}) \right\} \tag{12}$$

where μ^i is the intensity-average function of the ith colour channel.

The proposed segmentation can be performed at any colour space such as RGB or YUV. It has been recognised that selecting of an appropriate colour space produces more perceptually effective segmentation results [25, 26]. In particular, segmentation in YUV or LUV spaces often produces more favourable results than in RGB space [25–27]. Many of the images and image sequences in the databases are in YUV format where Y is in full resolution whereas the U and V components are in half resolution. The fact that the Y, U and V channels are presented at different resolutions is not considered in any of the existing regular single or multiresolution colour image segmentation algorithms. However, this fact calls for a specially fitted multiresolution algorithm to perform the segmentation task effectively. The proposed algorithm has enough flexibility to directly segment this format of colour images. It is possible that only the available components of colour data at different resolutions be considered to classify the vector of corresponding pixels at different resolutions under one of the segmentation labels. In other words, it is possible that the terms related to the chrominance components at the highest resolution be deleted from the objective function in (12). Tying the corresponding pixels together at different resolutions is equal to considering all the available colour information which classifies the vector and pixels at different resolutions successfully. Considering the same argument, the objective function can be simplified to segment the grey-level image.

3.4 Smoothness criterion

Object borders are one of the most important properties for visual perception. Many natural objects exhibit smooth borders/edges. Hence, to some extent there is a correlation between visually pleasing objects and edge/border smoothness. Psychologically, smoother edges/borders increase the perceived visual quality of the segmentation result. Therefore in some edge/contour-based segmentation algorithms such as the active contour model and the 'Canny' edge extraction algorithm, the extracted objects, regions, edges or borders are smoothed [28–30]. Some shortcoming of edge-based approaches for segmentation are unclosed contour detection, problems in the texture or noisy environments [28, 30], the need for initial estimation, detection of only one object in the scene, computational complexity and convergence problems in detecting convex regions [28] and so on. While most of these problems are overcome in region-based approaches, the smoothness criterion has not been used in region-based approaches yet.

Traditionally, in region-based image/video segmentation algorithms, the image features such as pixels' grey-levels or

colours have been considered. In most of these approaches, emphasis is put on the accuracy of segmentation. However, the shape delineation of objects/regions and producing a well-pleasing objects'/regions' shape have not attracted enough attention. On the other hand, perfect segmentation, if not impossible, is very difficult and distortions created by wrong segmentation in region-based approaches can result in incorrect, rough and unpleasing borders/edges. For example in pixel-wise segmentation algorithms such as MRF-based algorithms, the segmentation algorithm sometimes cannot capture the object/region structure very well, especially in low contrast areas which can result in border fluctuation. Therefore in the proposed region-based segmentation algorithm, a smoothness criterion is incorporated into the objective function, which improves the visual quality of the segmentation result.

Due to multiresolution object extraction applications such as scalable coding, the smoothness constraint is emphasised by considering it in the proposed multi-resolution scalable segmentation's analysis. At high resolutions, the large number of pixels ensures more visual quality for the segmentation. However, at lower-resolutions the visual quality can suffer due to insufficient information and downsampling distortion. Downsampling distorts shapes and cannot necessarily preserve their topology at lower-resolutions for all possible shapes [13]. This is more critical for complex shapes in terms of the ratio between perimeter and area pixels. Therefore achieving visually pleasing objects/regions at higher-resolutions does not necessarily ensure similar quality at lower-resolutions. Hence, it is necessary to enhance smoothness at all resolutions.

The proposed smoothness definition is based on the border's curvature, which is the rate of the angle change between a curve and the tangent line to the curve. In a digital environment, an estimation of curvature can be used. The estimation is explained in Fig. 4. Minimising the proposed estimation of smoothness prevents unwanted fluctuations in the border pixels.

Therefore the objective function is extended according to the following equation

$$E(X) = \sum_{\{S\}} \left\{ \sum_{i=1}^{3} \| Y_i(\{\boldsymbol{s}\}) - \mu^i_{X(\{\boldsymbol{s}\})}(\{\boldsymbol{s}\}) \|^2 + \sum_{\{\boldsymbol{r}\}\in\partial\{\boldsymbol{s}\}} V_c(\{\boldsymbol{s}\},\{\boldsymbol{r}\}) + \sum_{q\in\{\boldsymbol{s}\}} l_{\mathrm{res}(q)} * k(q) \right\} \quad (13)$$

where $k(q)$ shows the curvature estimation of pixel q, a pixel of vector $\{\boldsymbol{s}\}$, and $l_{\mathrm{res}(q)}$ is a coefficient which can be resolution dependent. To more emphasise on the lower resolution smoothness or visually pleasing $l_{\mathrm{res}(q)}$ increases when resolution decreases (the values of $l_{\mathrm{res}(q)}$ can be application dependent. In our examples, it is doubled when resolution is halved). In grey-level images, only the grey-level/intensity channel is available, and (13) can be reduced to

$$E(X) = \sum_{\{S\}} \left\{ [Y(\{\boldsymbol{s}\}) - \mu_{X(\{\boldsymbol{s}\})}(\{\boldsymbol{s}\})]^2 + \sum_{\{\boldsymbol{r}\}\in\partial\{\boldsymbol{s}\}} V_c(\{\boldsymbol{s}\},\{\boldsymbol{r}\}) + \sum_{q\in\{\boldsymbol{s}\}} l_{\mathrm{res}(q)} * k(q) \right\} \quad (14)$$

where Y is the grey-level/intensity function and μ is the grey-level/intensity average function.

The proposed smooth object extraction is different from a simple objects' border smoothness as has been done in the work of Marques and Llach [31] which is a filtering of the extracted video object shape to remove the small elongations introduced during the segmentation process. The differences are in the following areas: (a) our smoothing process takes part in the segmentation algorithm and changes the segmentation outcome; (b) with sufficient contrast, the proposed algorithm produces borders that are more faithful to the region's true shape; (c) on some occasions, some background pixels are added to the foreground regions to produce better looking shapes, especially at different resolutions; (d) by changing the smoothness coefficients (l_{res}(res)), the emphasis on the smoothness can be adjusted at different resolutions which produces visually pleasing shapes at different resolutions.

As an example of smoothness effect in spatial segmentation, consider the circle in Fig. 5*a*. It has two grey-levels, 100 in the background area and 200 in the foreground area. Considering a uniform noise in the range (0, 50) added to the background and subtracted from the object intensity. This noise changes the image from binary to grey-level and reduces the pixels intensity variation of the foreground to the background pixels. The image is segmented by the proposed algorithm at two resolutions 20 × 20 and 10 × 10. The lower-resolution smoothness is augmented by decreasing the smoothness coefficients to zero for the highest level and increasing the smoothness coefficient for

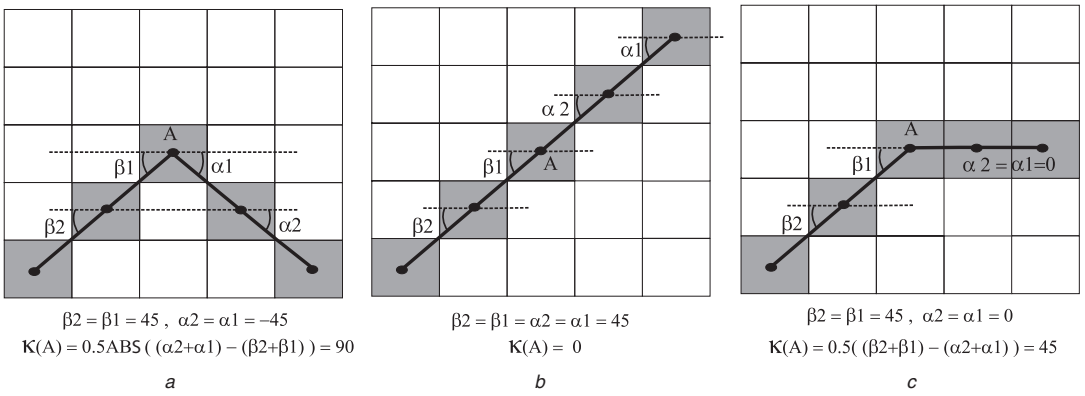

Fig. 4 *Curvature estimation*
a Corner point, $k = 90$
b Same direction $k = 0$
c Change direction point $k = 45$

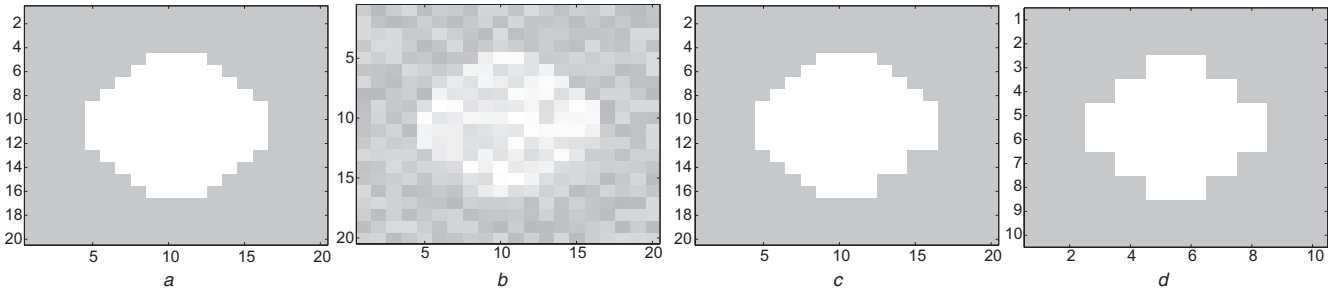

Fig. 5 *Scalable segmentation of a digital circle with emphasis on low-level smoothness*

a Original image
b Noisy image
c Segmentation at 20 × 20 resolution
d Segmentation at 10 × 10 resolution

lower-resolution to 10. The other parameters of segmentation are $k = 2$ and $\beta = 10$. The results are shown in Figs. 5*c* and *d* In this example, the smoothness criterion has deleted some pixels of the shapes at different resolution. The results could be compared with Figs. 1*a* and *b* at low and high resolutions considered as regular segmentation. The proposed segmentation method extracts a more pleasing shape at lower resolutions, albeit sometimes adding some distortion at higher resolution. However, larger number of pixels at higher resolutions increase visually pleasing effect of the extracted objects/regions.

3.5 MAP estimation

The iterated condition mode (ICM) optimisation method [16] is used to minimise the objective function in (13). The segmentation is initialised with the *k*-means clustering algorithm for each channel separately. Then neighbouring pixels with equal labels at all three channels form a region. The segmentation estimation is improved using ICM optimisation [16]. In single-resolution image segmentation, ICM optimises the objective function pixel by pixel in a raster-scan order until convergence is achieved. At each pixel, the segmentation of the processed pixel is updated given the current X at all other pixels. Therefore only the terms in the objective function related to the current pixel need to be minimised

$$E(X(s)) = \sum_{i=1}^{3} (Y_i(s) - \mu^i_{X(s)}(s))^2 - \sum_{r \in \partial_s} V_c(s, r) \qquad (15)$$

ICM was used in the single-level segmentation algorithm for grey-level images by Pappas [23] and was extended to colour images by Chang *et al.* [25]. In this paper, it is adapted to the multiresolution scalable segmentation algorithm. Similar to ICM optimisation technique for single-resolution image segmentation, the objective-function terms corresponding to the current vector are optimised given the segmentation at all other vectors of the pyramid. The resulting objective function term related to the current vector is

$$E(X\{\boldsymbol{s}\}) = \sum_{i=1}^{3} \|Y_i(\{\boldsymbol{s}\}) - \mu^i_{X(\{\boldsymbol{s}\})}(\{\boldsymbol{s}\})\|^2 + \sum_{\{\boldsymbol{r}\} \in \partial\{\boldsymbol{s}\}} V_c(\{\boldsymbol{s}\}, \{\boldsymbol{r}\}) + \sum_{q \in \{\boldsymbol{s}\}} l_q \cdot K(q) \qquad (16)$$

For grey-level images there is only the intensity channel and the objective function is simplified to

$$E(X\{\boldsymbol{s}\}) = (Y(\{\boldsymbol{s}\}) - \mu_{X(\{\boldsymbol{s}\})}(\{\boldsymbol{s}\}))^2 + \sum_{\{\boldsymbol{r}\} \in \partial\{\boldsymbol{s}\}} V_c(\{\boldsymbol{s}\}, \{\boldsymbol{r}\}) + \sum_{q \in \{\boldsymbol{s}\}} l_q \cdot K(q) \qquad (17)$$

During the optimisation process for each pixel $\boldsymbol{s}$ of a vector $\{\boldsymbol{s}\}$, the terms $\mu^i(s)$, $i = 1, 2, 3$, are estimated by averaging the channel intensities of all pixels that belong to the region i and are inside a window with width w centred at pixels s. The window size w is doubled when moves to the next higher-resolution. The average of any pixel s and its correspondences at all other levels in $\{\boldsymbol{s}\}$ are used to classify the pixels of $\{\boldsymbol{s}\}$ to a label which minimises (16). To reduce computational complexity, it is enough to consider only labels of $\{\boldsymbol{s}\}$ and its neighbouring vectors to select the best label by the energy minimisation through (16). Therefore for the pixels inside a region there is no computation and the regions' borders are gradually refined. Furthermore, this border processing prevents isolated noise pixels from becoming a new cluster, resulting in fewer wrongly detected boundaries [24].

Let us consider the overall optimisation algorithm now. As mentioned earlier, the initial segmentation of the pyramid is obtained by the *k*-means clustering algorithm [32, 33]. The pyramid's pixels are processed progressively from low to high resolutions. At each resolution, pixels are visited in a raster scan order. Intensity average $\mu^i(s)$, $i = 1, 2, 3$, at each pixel s and its corresponding pixels at the other resolutions for all possible classes are estimated with a pre-determined window size w used for estimation. Then the estimate of $X\{\boldsymbol{s}\}$ is updated using the ICM approach with a multi-level analysis using (17). By updating the segmentation labels of pixels at the current resolution, the corresponding pixels at the other levels are also updated. After convergence at the current resolution, the algorithm moves to the next higher resolution and updates the estimates of μ and X and so on, until all resolutions are processed. The stopping criterion at each resolution is the number of X updates which should be below a pre-defined threshold. Other optimisation and convergence criteria can also be used [21–23]. The whole procedure is repeated with a smaller window size. The algorithm stops when the predetermined minimum window size for the lowest resolution is reached.

4 Semantic video object extraction

At the core of semantic video segmentation is always a tracking algorithm. Tracking techniques are classified as forward or backward methods. In the forward method, the current frame objects/regions are projected to the next frame [7, 34, 35]. Conversely, in the backward method, the objects/regions are back projected to the previous frame using motion information. In forward tracking algorithms, the projected regions often need to be adjusted through a post-processing stage [34]. In backward tracking, the spatial segmentation gives the precise borders of objects [8]. This overcomes the problem of non-rigid moving objects. Therefore in this papers a multiresolution backward tracking algorithm is used. To detect the newly appeared objects/regions, the tracking algorithm is extended to present a video segmentation routine.

In the first frame, through user's intervention and spatial segmentation, a meaningful object (foreground) is determined. In special scenes such as 'head and shoulder' or 'car on the road', automatic image object extraction algorithms can be used [36, 37]. However, these algorithms are not matured yet, and they have many limitations such as detecting only predefined objects in a specific scene. High computational complexity is another problem of these algorithms. As a result, they cannot be used in a generic scene and need further development [38]. In the subsequent frames, the object (foreground) is tracked in an automatic procedure. In each frame, a multiresolution spatial segmentation is followed by an MRF-based backward region classification stage, which decides whether the regions belongs to the foreground or background. The images at different resolutions of the pyramid are separated to different regions by the proposed scalable image segmentation algorithm and the same segmentation patterns are produced at different resolutions. This feature is used in the proposed tracking algorithm. On the basis of the region size, each region is processed at the proper resolution and results are extended to corresponding regions in the other resolutions of the pyramid. Classifying larger regions at lower resolutions significantly reduces the computational complexity of the classification algorithm.

The proposed multiresolution video object extraction algorithm, with scalability, extends the attractive features of multiresolution image segmentation to video segmentation algorithm. Some of the improvements are better noise tolerance, faster classification and less computational complexity.

4.1 Objective function of MRF-based region classifier

MRF-based processing is the most frequently used stochastic model in image processing and computer vision. It has the ability to capture the spatial continuity of natural images, and similarly it can capture the spatial and temporal continuity of video signals. Pixel-based processing increases the computational complexity of the algorithm, and therefore in this work, MRF-based classification is used for region labelling. Regions are obtained from the scalable spatial segmentation; region-based processing therefore increases the spatial accuracy of the video segmentation processing. Since the number of regions is much lower than the number of pixels, the presented algorithm is very effective.

The proposed algorithm starts by partitioning the current frame into different regions using the proposed scalable spatial segmentation algorithm. Then by an MRF-based objective function, each region and its corresponding regions at the other resolutions are classified to foreground or background. At first the region classifier's objective function is extracted at single resolution and then it extends to multiresolution scalable mode.

According to the MAP estimation criterion, the conditional probability of video segmentation labelling X, given the observations, should be maximised. The observations include the last frame segmentation X^-, motion information θ and colour/intensity I of the current frame. Hence using the Bayes theorem, we obtain

$$P(X|X^-, \theta, I) \propto P(X^-|\theta, X, I)P(\theta|X, I)P(X|I) \tag{18}$$

The first term on the right-hand side of (18) explains the temporal continuity of the segmentation field. This term encourages corresponding regions/pixels at the subsequent frames to have the same label. Considering MRF-based modelling for the labelling process, the conditional probability of the estimated label field at the previous frame X^- is modelled as a Gibbs distribution

$$P(X^-|X, \theta, I) = \frac{1}{z_1}\exp\{-E_T(X, X^-, \theta, I)\} \tag{19}$$

where z_1 is a normalisation constant that does not affect the optimisation process. The energy term $E_T(X, X^-, \theta)$ is modelled by the Gibbs distribution potentials $V^T_{R_i}$ over single cliques consisting of just one region as follows

$$E_T(X^-, \theta, X, I) = \sum_{i=1}^{k} V^T_{R_i}(X^-, \theta, X, I)$$

$$V^T_{R_i}(X^-, \theta, X, I) = z_t Q(R_i) \tag{20}$$

In this equation, k is the number of regions, and index i points to different regions. z_t is a normalisation constant. $Q(R_i)$ is the number of pixels in R_i which after the back projection process have different labels compared to the current frame. Therefore a smaller Q indicates a higher probability for the region to have the same label as the corresponding projection at the previous frame determined by θ_{R_i}. The coefficient z_t determines the trend to track the same label field for corresponding regions in consecutive frames. This term also allows tracking of stationary objects/regions.

The second term on the right-hand side of (18) is a motion constraint which explains the relationship of the motion vectors to the labelling process. It is modelled as a Gibbs distribution

$$P(\theta|X, I) = \frac{1}{z_2}\exp\{-E_M(\theta, X, I)\} \tag{21}$$

where z_2 is a normalisation constant which does not affect the optimisation process. Considering the compensated, global motion and the labels set as F, B, the above-mentioned constraint for labels along the motion trajectory means that any non-zero motion vectors indicate foreground areas. Therefore the energy term is formed by the Gibbs potential function as

$$E_M(X, \theta, I) = \sum_{i=1}^{K} V^M_{R_i}(X, \theta, I) \tag{22}$$

where energy term $V^M_{R_i}$, corresponding to the region R_i, is described as follows

$$V^M_{R_i}(X, \theta, I) = \begin{cases} -\alpha A(R_i) & (X_{R_i} = F \text{ and } \theta_{R_i} \neq 0) \text{ or} \\ & (X_{R_i} = B \text{ and } \theta_{R_i} = 0) \\ +\alpha A(R_i) & (X_{R_i} = F \text{ and } \theta_{R_i} = 0) \text{ or} \\ & (X_{R_i} = B \text{ and } \theta_{R_i} \neq 0) \end{cases} \tag{23}$$

where $A(R_i)$ is the size of region R_i, and α is a coefficient. This term encourages moving regions to be classified as foreground. The magnitude of the motion vector is not considered, but only whether it is zero or not. Since this term detects the newly appeared/moved objects/regions, its effect is similar to that of a change detector.

The third term on the right-hand side of (18) models the spatial continuity of the segmentation field. It is modelled as a Gibbs distribution whose energy term E_{S_p} is formed by the Gibbs potentials V^{S_p} as a clique function of two neighbouring regions R_i and R_j as follows [39]

$$P(X|I) = \frac{1}{z_3}\exp\{-E_S(X,I)\}$$

$$E_S(X,I) = \sum_{i,j=1,i\neq j}^{k} V^{S_p}_{R_iR_j}(X,I)$$

$$V^{S_p}_{R_iR_j}(X,\theta) = \begin{cases} -z_f \cdot f(M_{R_i} - M_{R_j}) \cdot N_{R_iR_j}, & X_{R_i} = X_{R_j} = F \\ -z_b \cdot f(M_{R_i} - M_{R_j}) \cdot N_{R_iR_j}, & X_{R_i} = X_{R_j} = B \\ z_{\text{diff}} \cdot f(M_{R_i} - M_{R_j}) \cdot N_{R_iR_j}, & X_{R_i} \neq X_{R_j} \end{cases} \quad (24)$$

where z_3 is a normalisation constant, k is the number of regions, $N_{R_iR_j}$ is the length of the common border between regions R_i and R_j. M_{R_i} and M_{R_j} are the means of regions R_i and R_j, respectively. f is a function which compares regions' mean values and gives a small value for dissimilar regions and a large value for similar regions. A good definition for f is given by Tsaig and Averbuch [40], which is shown in Fig. 6. The corresponding formula can be expressed as

$$f(d) = \begin{cases} T_h & d < d_l \\ T_l - \dfrac{T_h - T_l}{d_h - d_l}(d - d_l) & d_l < d < d_h \\ T_l & d > d_h \end{cases} \quad (25)$$

where T_l, T_h, d_l and d_h are the entered thresholds. Therefore two regions with similar spatial properties are more likely to have the same label.

Since the classification function is modelled as an MRF processing, its probability function in (18) can be modelled as a Gibbs distribution as follows

$$P(X|X^-, \theta, I) = \frac{1}{z_L}\exp\{-E_L(X|X^-, \theta, I)\} = \frac{1}{z_L}\exp\left\{-\sum_{i=1}^{k} U_L(X_{R_i}|X^-, \theta, I)\right\} \quad (26)$$

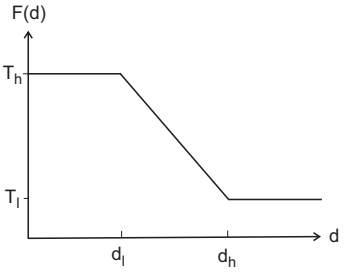

Fig. 6 *Similarity function [39]*

Therefore considering (18)–(22), (24) and (26), the objective function for the scalable multiresolution video segmentation is equal to using

$$E_L(X|I, X^-, \theta) = \sum_{i=1}^{K} \left\{ z_t \cdot Q(R_i) \pm \alpha \cdot A(R_i) + \sum_{P\in\partial R_i} z_x \cdot f(M(R_i) - M(P)) \cdot N_{R_iP} \right\} \quad (27)$$

where K is the number of regions, ∂R_i are the neighbouring regions of region R_i and finally Q, A and f are the functions as defined in (20), (23) and (25). The set of neighbouring regions of R_i is shown by ∂R_i, and the coefficient z_x comes from (24) and is equal to

$$z_x = \begin{cases} -z_f & X(R_i) = X(P) = F \\ -z_b & X(R_i) = X(P) = B \\ z_{\text{diff}} & X(R_i) \neq X(P) \end{cases}$$

4.2 Objective-function optimisation

The objective function should be optimised by one of the MRF optimisation methods. However, at first, an initial estimation is necessary. The initial estimation is obtained by considering the temporal continuity term. The regions are simply back projected to the previous frame, and the number of object pixels is counted. If the ratio of the counted object pixels over the area of a region exceeds a threshold, the processed region is considered as a foreground area. In multiresolution mode, the average of the computed ratio at different resolutions is compared with the threshold. Then an ICM-like optimisation is performed. However a raster scan of regions, unlike the raster scan of an image's pixels, does not have a physical interpretation. Since large-size regions are more likely to be classified correctly, regions are put in a queue in the order of their size from large to small regions. The correct classification of large regions can help with the right classification of their neighbouring small regions. Regions are visited according to the priority queue. For any region such as s, the terms of the objective function corresponding to this region are optimised given the classification of all the other regions. Hence, considering (27) and (28), the objective function related to region s is stated as follows

$$U_L(R_i) = z_t Q(R_i) \pm \alpha A(S) + \sum_{P\in\partial(R_i)} z_X f(M(R_i) - M(P)) N_{R_iP} \quad (28)$$

One cycle of optimisation process continues until the queue is empty. The convergence criterion updates more than a threshold value such as 5% of regions, in one cycle of region visits. To reduce the computational complexity, regions which, when back projected to the previous frame, are covered by foreground (background) pixels by more than a threshold (i.e. 90%) do not need reclassification, and they take part in the objective function only for classification of their neighbouring regions. The different coefficients are determined empirically.

However, more reduction in the computational complexity is achieved by classifying each region in a proper resolution and extending the result to the corresponding regions at the other resolutions. Depending on the size of the region and the defined thresholds, a resolution is selected, the

region at that single-resolution is classified, and the result is extended to the other lower- and higher-resolutions. For example the largest regions are classified at the lowest resolution, and very small regions are classified at the highest resolution. This significantly reduces the computational complexity because motion estimation and back projection to the lower-resolution has much less computational complexity than working at the higher-resolutions.

Since functions Q, A, f and N can be extended to multi-resolution mode, the proposed objective function for the video segmentation in (27) or (28) can be extended to multi-resolution mode. These functions are performed on the array of corresponding regions at different resolutions. The idea is that the single resolution function be computed over corresponding regions at different resolutions of the pyramid and then be averaged. Owing to the smaller size of lower resolutions, the result of lower resolutions (in case) can be scaled by the resolution's size reduction factor over pyramid such as 4, 16 or 64 to make the average more accurate. For example at low resolution phase of function Q, counting the number of pixels with different labels following region back projecting to the previous frame is first scaled by a proper factor and then it is averaged with the function values at other resolutions.

The proposed objective function does not need the exact motion vectors. Therefore a simple translational motion model in the following equation is used, which significantly reduces the computational complexity

$$\hat{\boldsymbol{v}}_x = b, \quad \hat{\boldsymbol{v}}_y = c \tag{29}$$

For small size regions, the assumption of constant motion vector is justified. In addition, exact motion compensation is not required at this stage, and classifying only the foreground region is enough. Motion estimation is obtained by shifting the region over the last frame and finding the best match in a hierarchical framework. The hierarchical search is started at the coarsest resolution and propagates to the higher resolutions while the motion estimation is refined at each resolution until fine resolution is achieved [40]. At each resolution, a threshold determines the area of the region search. The second energy term E_M in the objective function encourages regions with non-zero motions to be classified as foreground. The problem behind this classification is the occlusion related to covered and uncovered regions [41, 42]. Backward apparent motion classifies these regions as moving regions, and in the classification they might be incorrectly detected as foreground regions. To overcome this problem, only valid motion vectors in the energy term (E_M) in the objective function are processed. The backward motion vector such as ($\boldsymbol{v}_{x1}$, $\boldsymbol{v}_{y1}$) computed for region A is valid, if the corresponding forward motion vector from the projected regions in the previous frame towards current frame is in the opposite direction. However, in practice, some variations could be tolerated and a threshold for the differences can be determined. These will project the corresponding region in the previous frame to region A. Figure 7 explains this relationship. Otherwise, this motion vector is called invalid and is replaced with the zero vector. This replacement prevents the detection of uncovered regions.

4.3 Object's border fine tuning

For most of the object-based applications such as video editing and manipulation, the 'object-of-interest' should be extracted with pixel-wise accuracy. However, the proposed scalable grey-level segmentation can result in under-segmentation and may fail in discriminating between foreground and background objects in areas with low contrast. One way to increase the discriminating power of the segmentation is by using colour segmentation, which partitions the image into more regions than the grey-level segmentation. This decreases the under-segmentation, but increases the computational complexity of spatial segmentation. However, in some image sequences with low colour contrast, under-segmentation can still happen. In this case, the suggestion is to divide the image into watershed basins, which results in an over-segmentation including many small regions [43, 44]. The region growing algorithms can also produce over-segmentation, but the watershed is more faithful to the natural borders.

To retain the smoothness feature of the extracted regions and ensure visually pleasing segmentation, the scalable multiresolution grey-level/colour image segmentation is used. The regions which are smaller than a threshold are left, and the other regions are divided into smaller basin regions by the watershed algorithm [44]. The watershed basins are also downsampled to lower-resolutions to create the corresponding regions at the lower-resolutions. Subsequently, the vector basin regions are classified. This leads to avoiding the unnecessary partitioning of small regions and retaining most of the aesthetically pleasing borders resulting from the scalable segmentation. Fig. 8 shows the idea. For the spatial segmentations of the frame displayed in Fig. 8*a*, the partitioning of the regions to the basins is shown in Fig. 8*b*.

Partitioning into basins overcomes the under-segmentation problem, but it significantly increases the

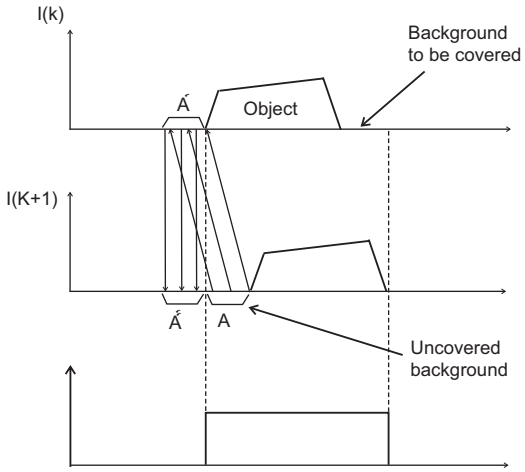

Fig. 7 *Detection of uncovered background [41]*
Region A' at frame k is projected to region A at frame $k+1$, but region A is back projected to region A'

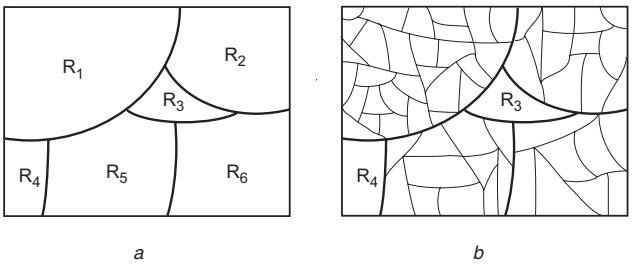

Fig. 8 *Partitioning the segmentation regions to the basins*
a Original (scalable) image segmentation
b Partitioning the segmentation regions to the basins
Regions R_2 and R_3 are smaller than the predefined threshold and have not been divided to basins

Table 1: Parameters for the proposed and alternative algorithms

Sequence	Spatial segmentation	Video segmentation algorithm									
	Number of clusters	β	z_t	z_f	z_b	z_{diff}	α	T_h	d_h	T_t	d_l
Claire	5	50	10	5	5	5	15	2	80	0.5	20
Table-Tennis	6	100	10	5	5	5	15	2	80	0.5	20
Hall-Monitor	7	50	10	5	5	5	15	2	80	0.5	20
Mother and Daughter	7, 2, 2	40	10	5	5	5	15	2	80	0.5	20

number of regions and the computational complexity of the labelling optimisation process. In addition, due to having more information in the large-size regions, their classification is also more confident than for small-size basin regions. However, the challenge is how to automatically determine the use of grey-level or colour segmentation and whether the partitioning of the image into watershed basin regions is necessary or not. It is clear that it depends on the contrast between foreground and background. However, except through human intervention, we are not aware of any effective solution for an automatic decision to choose regular or over-segmentation for generic application. This is somewhat similar to the problem of threshold and parameter tuning that requires many thresholds and parameters to be set by the users in different algorithms for image/video processing and generally in signal processing algorithms.

5 Experimental results and discussion

To evaluate the performance of the presented algorithm, four different sequences, 'Claire', 'Hall-Monitor' with CIF format, 'Table-Tennis' with SIF format and colour sequence 'Mother and Daughter' with QCIF format are segmented. The simulations were performed on a Pentium 4 PC computer with 2.4 GHz CPU clock and 512 Mbytes RAM. The algorithms were coded in the Microsoft Visual C++ 6.0 environment and Matlab software was also used for user interface and input/output functions. The parameters for the spatial segmentation and video object extraction algorithms are shown in Table 1. These parameters are set empirically. Further research is needed for their automatic tuning.

At the initial step, the user determines the rough boundary of the 'object-of-interest' through a graphic user interface (GUI). Subsequently, all regions with the majority of their area, more than a predetermined percentage (i.e. 50%), located inside this closed contour are selected to belong to the extracted object. This is more explained in the first example of this section.

In the first example, the proposed video segmentation algorithm is run over the 75 frames of the sequence. The user's selection of the 'Claire' object in the first frame of the 'Claire' sequence is shown in Fig. 9*a*. The subsequent

Fig. 9 *First frame 'Claire' object separation*

a Rough object separation by user intervention
b Segmentation by the proposed algorithm
c Extracted object at the first frame
d Object extracted in frame 20
e Object extracted in frame 45
f Object extracted in frame 65

Table 2: 'Claire' sequence smoothness

	88 × 72	144 × 176	288 × 352
Scalable tracking	54.67	54.7	53.15
Regular tracking	58.95	58	56.87
Improvement, %	7.54	6.03	6.77

spatial segmentation by the proposed algorithms is shown in Fig. 9*b*. The exact borders of the object at finest resolution are shown in Fig. 9*c*. The extracted objects at frames 20, 40 and 60 in multiresolution mode are shown in Figs. 9*d–f*.

To compare the proposed algorithm with other region-based object tracking and extraction methods, an alternative tracking algorithm is used. It is an ordinary backward tracking algorithm [45, 46] which includes only the temporal continuity term at the highest resolution. First, the current frame is partitioned into different regions by the MRF-based single resolution image segmentation proposed by Pappas [23]. Each region is then back-projected to the previous frame. If the number of projected pixels inside the foreground area at the previous frame is more than a threshold (i.e. 50% of the region's area), the region is classified as a foreground region. The alternative algorithm will be called the 'regular (backward) tracking algorithm'. The quantitative criterion for comparing the extracted objects by different algorithms is border smoothness which is averaged over the curvature of object border pixels. Although it is not an ideal criterion, it is well consisted with the results of our subjective tests. The smoothness comparison for the 'Claire' sequence for the three resolution levels are shown in Table 2. (The proposed scalable tracking algorithm directly produces the object at different resolutions, however, the object produced by regular tracking algorithm is down-sampled to lower resolutions.) The smoothness term modifies the segmentation in areas of the image that have lower grey-level contrast. In the 'Claire' sequences, the regions around the head have lower contrast compared to the shoulder and body areas. If the only head area be considered, the smoothness improves by 13.17%, 11.5% and 10.5% at different resolutions. As a subjective test example, Fig. 10 shows the extracted objects of the 23rd frame of the 'Claire' sequence when using the scalable and a regular algorithm, respectively. In this figure, images of different resolutions are shown at the same size to highlight the details. The analysis of both images (by any viewer) clearly shows that our algorithm has extracted a smoother and more visually pleasing 'Claire' object.

In the second example, the standard MPEG-4 'Table-Tennis' sequence which has textured background with fast moving objects is processed. In Fig. 11, frames 10, 23 and 32 with the extracted objects by the proposed video segmentation algorithm are shown. For the comparison purpose, observe the extracted objects in frame 10 of the 'Table-Tennis' sequence that were extracted by the proposed scalable video segmentation algorithm and by the alternative algorithm which is single-level version of the proposed tracking algorithm without any smoothness criterion called as regular algorithm. The extracted objects by the proposed and regular video segmentation algorithms at three different resolutions are shown in Fig. 12. Subjective comparison shows the better visual quality of the object's extracted by the proposed object segmentation algorithm. For a quantitative comparison, the object smoothness for the first 35 frames of the sequence is measured which presented in Table 3. Again, if only the hand and fingers with the racket are considered, the smoothness is nearly doubled. The computational complexity of the multiresolution tracking algorithm is reduced typically to $<30\%$ of tracking at the finest resolution, because smaller regions and less motion decrease the complexity of the matching procedure at lower-resolutions.

The proven high noise tolerance of the multiresolution image segmentation [15] is extended to video segmentation by the proposed algorithm. (Noise sources can be from

Fig. 10 *'Claire' object 23rd frame*

a Scalable 288 × 352
b Scalable 144 × 176
c Scalable 72 × 88
d Regular 288 × 352
e Regular 144 × 176
f Regular 72 × 88

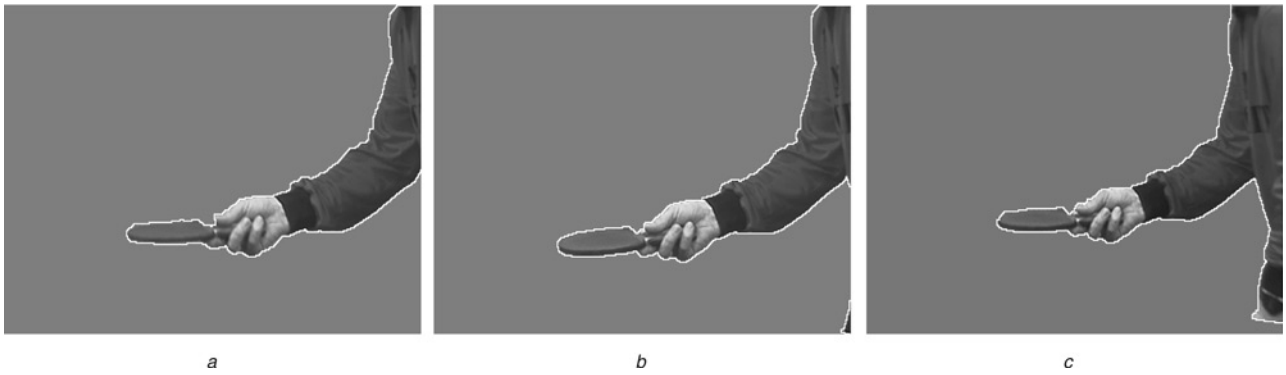

Fig. 11 *'Table-Tennis' object extraction*

a Frame 10
b Frame 23
c Frame 32

transmission channels, image acquisition or saving equipments, and so on.) In video object extraction, especially at low contrast areas, noise can adversely affect the regions matching, resulting in wrong classifications. For example some small background regions close to object areas are merged with the object and some regions belonging to the object areas are merged with the background. To overcome these matching errors, the proposed algorithm effectively uses the noise-reduced, lower resolution information to classify the regions. This is possible due to the proposed multiresolution video segmentation algorithm.

To test the algorithm in noisy environments, a uniform noise in the range $(-25, +25)$ is added to the 'Table-Tennis' sequence. The noisy sequence is segmented with the proposed algorithm and the results are compared with the single-level tracking algorithm. Table 4 presents the smoothness of both algorithms. The misclassified numbers of pixels for different resolutions are counted in Table 5. The number of misclassified object pixels in the scalable multiresolution video segmentation algorithm decreases to $\simeq$50% of the pixel misclassification of the regular single-level segmentation algorithm. This confirms the superiority of the multiresolution algorithm. Fig. 13 shows the extracted objects in frame 14 for both multiresolution and single-level object extraction.

In the third example, the 'Hall-Monitor' CIF sequence is segmented. In this example, the 'object-of-interest' appears gradually. Consequently, the change detector

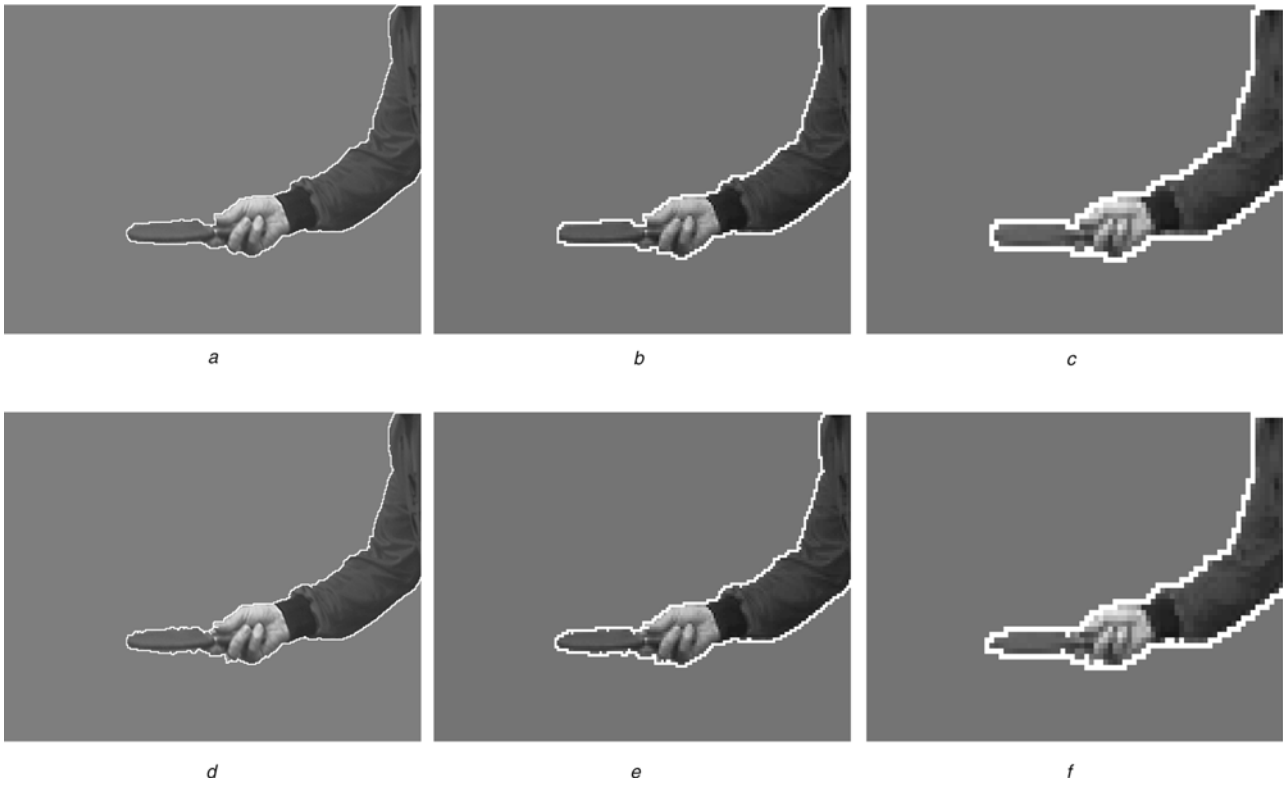

Fig. 12 *'Table-Tennis' object 10th frame*

a Scalable 240 × 352
b Scalable 120 × 176
c Scalable 60 × 88
d Regular 240 × 352
e Regular 120 × 176
f Regular 60 × 88

Table 3: 'Table-Tennis' sequence smoothness

	60 × 88	120 × 176	240 × 352
Scalable tracking	55.6	53.87	53.10
Regular tracking	58.82	57.63	56.22
Improvement, %	6.84	6.97	5.88

Table 4: Noisy 'Table-Tennis' smoothness

	60 × 88	120 × 176	240 × 352
Scalable tracking	56.73	55.42	55.55
Regular tracking	62.8	62.66	63.62
Improvement, %	10.7	13.1	14.54

Table 5: Misclassified object's pixels in noisy 'Table-Tennis'

	60 × 88	120 × 176	240 × 352
Scalable tracking	17	63	262
Regular tracking	35	134	528
Improvement, %	51	53	50

embedded in the second term of the MRF objective function identifies newly appearing objects/regions, whereas the tracking algorithm inherited in the first term of the MRF objective function detects already-present objects/regions. In this algorithm, due to low contrast of the foreground and background, the spatial segmentation cannot discriminate between the foreground and background in some areas of the image. Therefore the algorithm partitions the regions bigger than 20 pixels by the watershed algorithm, and the basin regions are classified.

The object of frame 40 extracted by the scalable algorithm at different resolutions is shown in Fig. 14. The extracted objects of frames 34, 44 and 60 using the scalable and the regular algorithms can be seen in Fig. 15. Some regions related to shaded areas are also detected as objects, because the shading between two consecutive frames is also changed.

Increasing the value of change-detector thresholds can reduce the size of detected areas of shading but increases the risk of missing some parts of the object during the detection process. As a subjective test, the comparison of the extracted objects in Fig. 15 confirms the superiority of the proposed video object extraction algorithm over the regular object-detection algorithm in creating a visually more pleasing segmentation. Table 6 confirms the improved smoothness of the proposed algorithm.

In the fourth example, the 75 frames of the QCIF size 'Mother and Daughter' colour image sequence are processed. The frames are in YUV format, where Y is in full resolution and U and V are in half resolution. At the first step, the grey-level image is segmented by the proposed scalable image segmentation algorithm. As Figs. 16*a* and *b* show, the spatial segmentation does not separate the object from background. In particular, the foreground and background regions around child's face and neck are mixed together. Colour information increases the discrimination and separation capabilities of the segmentation process. Therefore each frame of the 'Mother and Daughter' sequence is segmented by the proposed scalable colour image segmentation at three different resolutions. Figs. 16*c* and *d* show that the foreground regions are successfully separated from background and the next stages of the video segmentation algorithm can be performed. The 'object-of-interest' is selected by user intervention at the first frame, and it is tracked in the next frames by the proposed video-segmentation algorithm. In Fig. 17, frames 32, 50 and 68 are shown with the extracted objects at the highest resolutions.

In Fig. 18, the objects of frames 48, 58 and 72 extracted by the proposed scalable algorithm and regular backward tracking algorithm are compared. The objects extracted by the proposed object extraction algorithm are shown in the top row of the figure. The objects extracted by the regular tracking algorithm are shown in the second row. Subjective comparison of the extracted objects clearly shows better visual quality of the objects extracted by the proposed object segmentation algorithm.

The simulation details include the number of frames, size of frames, grey-level or colour images, with/without global motion estimation and compensation, divided to basins or not, average the time of frame processing and the number of processed frames per minute for the proposed scalable algorithm. Details of the proposed and the alternative algorithms for different sequences are shown in Tables 7 and 8. The strings '++' and '−' declare that the sub-process determined at that column's title is performed for that sequence or not. The following comparisons were made:

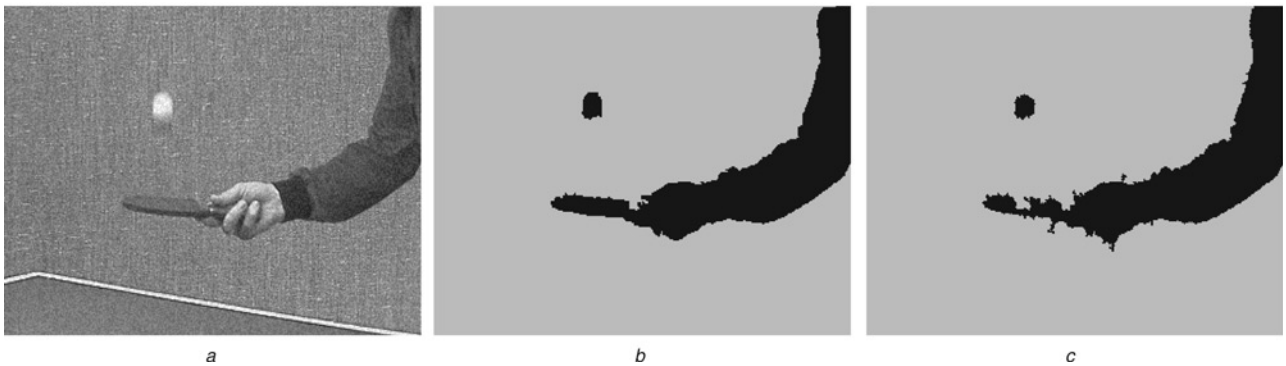

Fig. 13 *Object extraction from noisy 'Table-Tennis' sequence*
a Frame 14 at resolution 240 × 352
b Scalable object extraction at resolution 240 × 352
c Single-level object extraction at resolution 240 × 352

Fig. 14 *'Hall-Monitor' sequence object extraction at frame 40*
a Resolution 288 × 352
b Resolution 144 × 176
c Resolution 72 × 86

- The 'Claire' and 'Mother and Daughter' sequences: compared with regular backward tracking without global motion compensation.
- The 'Table-Tennis' and 'Hall-Monitor' sequences: compared with an algorithm similar to the proposed algorithm, but in the single resolution mode without the smoothness constraint.

The running times for the proposed scalable and the alternative algorithms are compared in Table 7. The alternative regular backward tracking algorithm is faster than the proposed algorithm. The main reason is the computation of the smoothness term. If the smoothness term is deleted from the scalable segmentation process, the computational complexity of the proposed scalable segmentation algorithm decreases to less than one third.

Although inherently the algorithm can be performed in real time, practically, as Table 8 shows, due to too much computational complexity the algorithms are not real time. In sequences such as 'Table-Tennis' which need global motion compensation, the computational complexity is much higher. Also, switching from the grey-level to colour segmentation nearly doubles the complexity. Similarly, decomposition of the segmented grey regions to basins increases the computational complexity by about three times. In some tracking algorithm such as that in the work of Zhou *et al.* [47], the global motion estimation is deleted, which decreases the computational complexity. However, this algorithm tracks the already-detected objects, and detecting newly appearing objects is not considered.

All the performed subjective and objective tests at this section confirm the superiority of the proposed video segmentation algorithm particularly in terms of visual quality of the extracted objects. However, for a more commonly used objective comparison, the results were also evaluated by a method used by MPEG standard [48, 49]. This

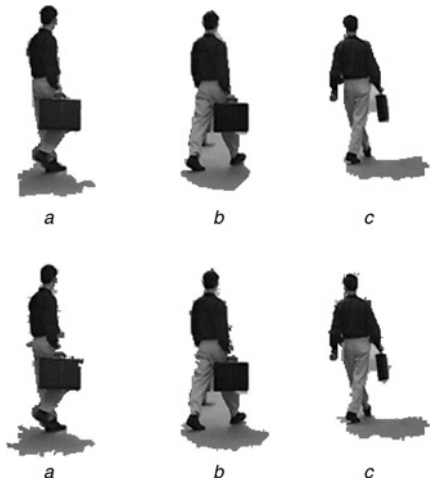

Fig. 15 *'Hall-Monitor' sequence object extraction*
a Scalable extraction at frame 34
b Scalable extraction at frame 44
c Scalable extraction at frame 60
d Regular extraction at frame 34
e Regular extraction at frame 44
f Regular extraction at frame 60

Table 6: 'Hall-Monitor' smoothness

	72 × 88	144 × 176	288 × 352
Scalable tracking	45.4	45	45.5
Regular tracking	54.9	56.8	53.6
Improvement, %	17.3	20.8	15.1

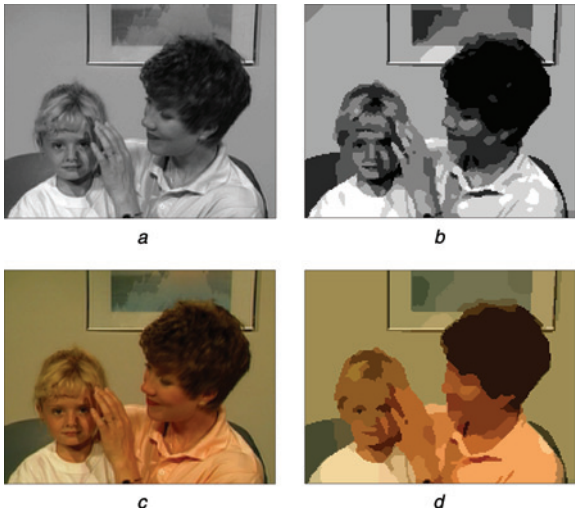

Fig. 16 *Frame 34 of 'Mother and Daughter' QCIF sequence segmentation with k = 7, 2, 2 clusters and β = 40*
a Original grey-level image
b Regular grey-level single resolution segmentation
c Colour image of 'Mother and Daughter' where U and V are in half resolution
d Proposed scalable segmentation

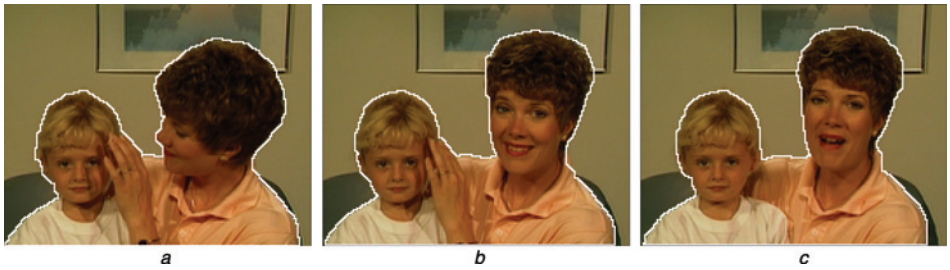

Fig. 17 *'Mother and Daughter' sequence object extraction*

a At frame 32
b At frame 50
c At frame 68

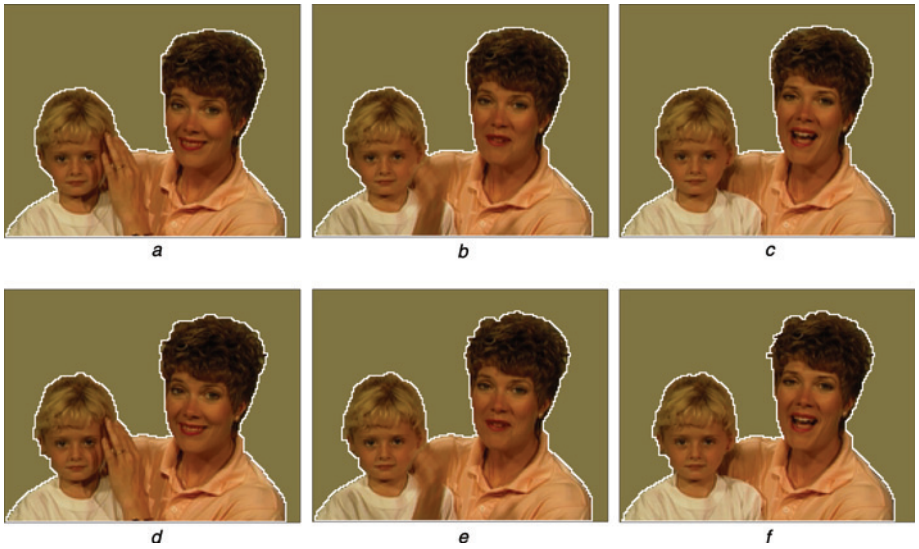

Fig. 18 *'Mother and Daughter' sequence object extraction by the proposed scalable algorithm and regular tracking algorithm at different frames*

a Scalable at frame 48
b Scalable at frame 58
c Scalable at frame 72
d Regular tracking algorithm at frame 48
e Regular tracking algorithm at frame 58
f Regular tracking algorithm at frame 72

method uses a reference segmentation (ground truth) to determine if each pixel is classified correctly or not. The ratio of misclassified pixels to the number of pixels determines the spatial segmentation accuracy for each object [48, 49]. For the first 25 frames of the four tested sequences in this section, the ground truth are extracted with pixel-wise accuracy. A graphical software lets the user determine the border of the interested objects manually with pixel-wise accuracy. The spatial accuracy is evaluated for both the objects extracted by the proposed algorithm and the regular algorithm used in each experiment. The results are shown in Table 9.

As the table shows, for the sequences with enough contrast between object and background and small object motion, the spatial accuracy for the proposed algorithm is close or a little less than the compared traditional algorithms. In these cases, most object extraction algorithms extract the objects. However, the smoothness criterion at the proposed algorithm misclassifies small number of

Table 7: Details of the proposed scalable video segmentation algorithm

Sequence	No. of frames	Size	Grey/ colour	Global motion	Basins
Claire	78	CIF	Grey	−−	−−
Table-Tennis	35	SIF	Grey	++	−−
Hall-Monitor	65	CIF	Grey	−−	++
Mother and Daughter	75	QCIF	Colour	−−	−−

Table 8: Running times of the proposed and alternative algorithms, performed on a 2.4 GHz Pentium 4

Sequence	Proposed scalable		Alternative algorithm	
	Sec/ frame	Frame/ min	Sec/ frame	Frame/ min
Claire	6.9	9	3.48	17
Table-Tennis	76	0.8	54.5	1.1
Hall-Monitor	19.3	3	13.92	4.3
Mother and Daughter	12.8	4.7	6.97	8.6

Table 9: Spatial accuracy for the extracted objects by the proposed and compared algorithms

Sequence	Claire, %	Table-Tennis, %	Hall-Monitor, %	Mother and Daughter, %
Proposed algorithm	99.8	98.5	87.4	99.4
Traditional algorithm	99.9	99.6	82.3	98.3

pixels around the object's contour to improve the visual quality of the extracted object. This introduces a small controlled error to improve the performance of the proposed algorithm according to the visual quality criterion. For more complex sequences with less contrast between object and background and fast moving objects, the multiresolution feature increases the spatial accuracy of the proposed algorithm compared to single-resolution algorithms in the literature. In this case, the error resulted due to smoothness criterion can be negligible compared to the errors resulted from other algorithms not suitable to cope with these complexities.

6 Conclusions

In this paper, a new semi-automatic MRF-based multiresolution video segmentation algorithm for VOP extraction is proposed. The objective function of the algorithm includes spatial and temporal continuity. Temporal continuity tracks the objects already extracted in the previous frames even when they stop. The motion constraint term detects newly appearing objects/regions. The motion-validity examination overcomes the occlusion problem. Region continuity considers the spatial consistency of the labelling algorithm. Region smoothness is introduced as a new criterion for region classification and is added to the objective function. The algorithm is extended to multiresolution by considering the corresponding regions at different resolutions and processing them in multidimensional or vector space. The final solution is obtained by the MAP criterion and an ICM-like optimisation method. The objects are extracted at different resolutions of the pyramid. The algorithm includes a version for object extraction from scenes with low grey-level or colour contrast. This version divides the region into watershed basin regions and classifies the basins. The proposed method provides fine localisation of the borders of regions. Multiresolution processing allows larger motion, better noise tolerance and less computational complexity. The algorithm also deals with the occlusion problem and corrects motion estimation. Comparison with different algorithms confirms the superiority of the proposed algorithm.

For further improvement of the algorithm, a more sophisticated solution for the occlusion problem can be considered. Better processing of the motion information to prevent shade detection is also necessary. Discrimination between different objects in the scene can be considered. More research is needed to determine the necessity of partitioning the segmentation into basins. Most of the computational complexity of the algorithm lies within the global motion estimation. Therefore more effective global motion estimation or deleting its role from the algorithm can be considered. Finally, more research into fully automatic object extraction including the identification of the 'object-of-interest' in the first frame and the automatic determination of the parameters and thresholds are necessary.

7 References

1 Pereira, F., and Ebrahimi, T.: 'The MPEG-4 book' (Prentice-Hall PTR, Upper Saddle River, NJ, 2002)
2 Manjunath, B.S., Salembier, P., and Sikora, T.: 'Introduction to MPEG-7: multimedia content description interface' (John Wiley & Sons, 2002)
3 Bertero, M., Poggio, T.A., and Torre, V.: 'Ill-posed problems in early vision', *Proc. IEEE*, 1998, **76**, (8), pp. 869–889
4 Tsaig, Y., and Averbuch, A.: 'Automatic segmentation of moving objects in video sequences: a region labeling approach', *IEEE Trans. Circuits Syst. Video Technol.*, 2002, **12**, (7), pp. 597–612
5 Meier, T., and Ngan, K.N.: 'Automatic segmentation of moving objects for video object plane generation', *IEEE Trans. Circuits Syst. Video Technol.*, 1998, **8**, (5), pp. 525–538
6 Cavallaro, A., and Ebrahimi, T.: 'Accurate video object segmentation through change detection'. Proc. IEEE Int. Conf. Multimedia and Expo (Cat. No.02TH8604), 2002, pp. 445–448
7 Kim, M., Choi, J.G., Kim, D., Lee, H., Lee, M.H., Ahn, C., and Ho, Y.-S.: 'A VOP generation tool: automatic segmentation of moving objects in image sequences based on spatiotemporal information', *IEEE Trans. Circuits Syst. Video Technol.,*, 1999, **9**, (8), pp. 144–50
8 Kim, Y.R., Kim, J.H., Kim, Y., and Ko, S.J.: 'Semiautomatic segmentation using spatio-temporal gradual region merging for MPEG-4', *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, 2003, **E86-A**, (4), pp. 2526–2534
9 Danyali, H., and Mertins, A.: 'Flexible, highly scalable, object-based wavelet image compression algorithm for network applications', *IEE Proc., Vis., Image Signal Process.*, 2004, **151**, pp. 498–510
10 Ta Hsiang, S.: 'Highly scalable subband/wavelet image and video coding'. PhD thesis, Rensselaer Polytechnic Institute, Troy, New York, 2002
11 ISO/IEC JTC1/SC29/EG11/N2322: 'MPEG-4 applications', July 1998
12 Danyali, H.: 'Highly scalable wavelet image and video coding for transmission over heterogeneous networks'. PhD thesis, School of Electrical, Computer and Telecommunication Engineering, University of Wollongong, Wollongong, NSW, Australia, 2003
13 Borgefors, G., Ramella, G., Sanniti di Baja, G., and Svenson, S.: 'On the multiscale representation of 2D and 3D shapes', *Graph. Models Image Process.*, 1999, **61**, (1), pp. 44–62
14 Tab, F.A., Naghdy, G., and Mertins, A.: 'Multiresolution image segmnetation for scalable object-based wavelet coding'. Proc. 7th Int. Symp. DSP for Communication Systems (DSPCS'03), 2003, Coolangatta, Qld, Australia, p. 171176
15 Tab, F.A., Naghdy, G., and Mertins, A.: 'Multiresolution image segmentation with border smoothness for scalable object-based wavelet coding'. Proc. 7th Int. Conf. Digital Image Computing Techniques and Applications (DICTA), Sydney, Australia, 2003, pp. 977–986
16 Besag, J.: 'On the statistical analysis of lattice systems', *J. R. Stat. Soc., Ser. B*, 1986, **48**, (3), pp. 259–279
17 Christopoulos, C., Skodras, A., and Ebrahimi, T.: 'The JPEG2000 still image coding system: an overview', *IEEE Trans. Consum. Electron.*, 2000, **46**, (4), pp. 1103–1127
18 Mertins, A., and Singh, S.: 'Embedded wavelet coding of arbitrary shaped objects', *Proc. SPIE*, 2000, **4067**, pp. 357–367
19 Li, S., and Li, W.: 'Shape-adaptive discrete wavelet transforms for arbitrarily shaped visual object coding', *IEEE Trans. Circuits Syst. Video Technol.*, 2000, **10**, (5), pp. 725–743
20 Xing, G., Li, J., Li, S., and Zhan, Y.: 'Arbitrarily shaped video-object coding by wavelet', *IEEE Trans. Circuits Systems Video Technol.*, 2001, **11**, pp. 1135–1139
21 Li, S.Z.: 'Markov random field modeling in image analysis' (Springer Verlag, Tokyo, Japan, 2001, 2nd edn.)
22 Tekalp, A.M.: 'Digital video processing' (Prentice-Hall, USA, 1995, 2nd edn.)
23 Pappas, T.N.: 'An adaptive clustering algorithm for image segmentation', *IEEE Trans. Image Process.*, 1992, **40**, (4), pp. 901–914
24 Meier, T., Ngan, K.N., and Grebbin, G.: 'A robust Markovian segmentation based on highest confidence first (HCF)'. IEEE Int. Conf. Image Process., 1997, vol. 1, pp. 216–219
25 Chang, M.M., Sezan, M.I., and Tekalp, A.M.: 'Adaptive Bayesian segmentation of color images', *J. Electron. Imaging*, 1994, **3**, (4), pp. 404–414
26 Luo, J., Gray, R.T., and Lee, H.C.: 'Towards physics-based segmentation of photographic color images'. Proc. Int. Conf. Image Processing, 1997, vol. 3, pp. 58–61

27 Gao, J., Zhang, J., and Fleming, M.G.: 'Novel technique for multiresolution color image segmentation', *Opt. Eng.*, 2002, **41**, pp. 608–614
28 Kass, M., Witkin, A., and Terzopoulos, D.: 'Snakes: active contour models', *Int. J. Comput. Vis.*, 1987, **1**, (4), pp. 321–331
29 Blake, A., and Isard, M.: 'Active contours: the application of techniques from graphics, vision, control theory and statistics to visual tracking of shapes in motion' (Springer, 1998)
30 Canny, J.: 'A computational approach to edge detection', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1986, **8**, (6), pp. 679–998
31 Marques, F., and Llach, J.: 'Tracking of generic objects for video object generation'. Int. Conf. Image Process. (ICIP), 1998, vol. 3, pp. 628–632
32 Ray, S., and Turi, R.: 'Determination of number of clusters in K-means clustering and application in colour image segmentation'. Proc. 4th Int. Conf. Advances in Pattern Recognition and Digital Techniques, 1999, pp. 137–143
33 Rosenberger, C., and Chehdi, K.: 'Unsupervised clustering method with optimal estimation of the number of clusters: application to image segmentation'. Proc. 15th Int. Conf. Pattern Recognition, 2000, vol. 1, pp. 656–659
34 Wang, D.: 'Unsupervised video segmentation based on watersheds and temporal tracking', *IEEE Trans. Circuits Syst. Video Technol.*, 1998, **8**, (5), pp. 539–546
35 Gu, C., and Lee, M.-C.: 'Semiautomatic segmentation and tracking of semantic video objects', *IEEE Trans. Circuits Syst. Video Technol.*, 1998, **8**, (5), pp. 572–584
36 Fan, J., Yau, D.K.Y., Elmagarmid, A.K., and Aref, W.G.: 'Automatic image segmentation by integrating color-edge extraction and seeded region growing', *IEEE Trans. Image Process.*, **10**, pp. 1454–1466
37 Tan, T.N., and Baker, K.D.: 'Efficient image gradient based vehicle localization', *IEEE Trans. Image Process.*, 2000, **9**, (8), pp. 1343–1356
38 Tab, F.A., and Naghdy, G.: 'Scalable multiresolution image segmentation and its application in video object extraction algorithm'. Proc. IEEE Int. Region 10 Conf. (TENCON), Melbourne, Australia, 2005
39 Tsaig, Y., and Averbuch, A.: 'Automatic segmentation of moving objects in video sequences: a region labeling approach', *IEEE Trans. Circuits Syst. Video Technol.*, 2002, **12**, (7), pp. 597–612
40 Chalidabhongse, J., and Kuo, J.: 'Fast motion vector estimation using multiresolution-spatio-temporal correlations', *IEEE Trans. Circuits Syst. Video Technol.*, 1997, **7**, (3), pp. 477–488
41 Tekalp, A.M.: 'Digital video processing' (Prentice-Hall, Upper Saddle River, NJ, USA, 1995, 2nd edn.)
42 Ngan, K.N., Meier, T., and Chai, D.: 'Advanced video coding principles and techniques' (Elsevier, Amsterdam, The Netherlands, 1999)
43 Patras, I., Hendriks, E.A., and Lagendijk, R.L.: 'Video segmentation by map labeling of watershed segments', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001, **33**, (2), pp. 326–32
44 Vincent, L., and Soille, P.: 'Watershed in digital space: an efficient algorithm-based on immersion simulations', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1991, **13**, (6), pp. 583–598
45 Gu, C., and Lee, M.: 'Semantic video object tracking using region-based classification'. Proc. Int. Conf. Image Processing, 1998, vol. 3, pp. 643–647
46 Perez, D.G., Sun, M.T., and Chuang, G.: 'Semiautomatic video object generation using multivalued watershed and partition lattice operators'. Proc. 2000 IEEE Int. Symp. Circuits and Systems, ISCAS 2000, Geneva, 2000, vol. 1, pp. 32–35
47 Zhou, J.Y., Ong, E.P., and Ko, C.C.: 'Video object segmentation and tracking for content-based video coding'. IEEE Int. Conf. Multimedia and Expo (ICME 2000), 2000, vol. 3, pp. 1555–1558
48 Villegas, P., and Marichal, X.: 'Perceptually-weighted evaluation criteria for segmentation masks in video sequences', *IEEE Trans. Image Process.*, 2004, **13**, (8), pp. 1092–1103
49 Wollborn, M., and Mech, R.: 'Procedure for objective evaluation of VOP generation algorithms'. MPEG Committee Document ISO/IEC JTC1/SC29/WG11 M2704, 1997