WELL DEFINED VIDEO OBJECT EXTRACTION SUITABLE FOR SCALABLE WAVELET BASED OBJECT CODING

Fardin Akhlaghian Tab, Golshah Naghdy* and Alfred Mertins***

*School of Electrical, Computer and Telecommunications Engineering
University of Wollongong, Wollongong, NSW 2522, Australia
Emails: {fat98 and golshah }@uow.edu.au

** Signal Processing Group, Institute of Physics, University of Oldenburg, Oldenburg, Germany
Email: alfred.mertins@uni-oldenburg.de

ABSTRACT

In this paper we present a semi-automatic multi resolution video objects extraction and tracking algorithm well suited to scalable wavelet based object coding. Objects of interest are determined in the first frame through initial user intervention followed by a spatial segmentation algorithm. The specified objects are afterwards tracked in the subsequent frames. The tracking algorithm includes Multiresolution Markov Random Field (MMRF) based spatial segmentation with emphasis on border smoothness in different resolutions and multi resolution backward partition projection. An intensity change detector indicates newly appeared objects/regions. The proposed method produces well defined and visually pleasing objects as well as allowing for larger motion tracking, better noise tolerance and less computational complexity.

1. INTRODUCTION

The vast demand for multimedia applications, stirred up by explosive growth in networking technology, especially the Internet, has led to rapid expansion in digital signal processing research in particular image coding and manipulation. Traditional block based image coding schemes suffer from great limitations, and as a result, new coding algorithms have emerged moving away from block based towards object based routines. This concept results in improved flexibility and interactive functionality. Video object or so called video object plane (VOP) extraction is a key issue in effectively applying the content based functionality. Furthermore, image/video segmentation is important in other applications such as machine vision and pattern recognition. Due to an increase in applications, a large number of automatic or semi automatic video object segmentation methods have been proposed [1], [2], [3]. However none of them consider multi resolutions object presentation, necessary for (spatial) scalability [4], and seldom has a multiresolution approach been proposed [5].

In a network environment such as the Internet, it is desirable that a large number of users with different processing capabilities and network bandwidth could access and transfer data easily. In such a heterogenous environment, a scalable coding 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. In scalable coding, the bitstream for low-end users is embedded as a subsets of the bitstream 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 [6]. Some of the desirable scalable functionalities are signal to noise ratio (SNR) scalability, spatial scalability and temporal scalability. Spatial scalability is a feature in the encoded bitstream that allows decoders to decode the image/video with different spatial resolutions. Therefore multi resolution VOP extraction is also a key issue in object based (spatially) scalable coding. On the other hand, because of attractive features of wavelet transform such as the potential to support SNR, spatial and temporal salabilities, wavelet based image/video coding have become increasingly important and have gained widespread acceptance. An example is the new JPEG 2000 still image compression standard [7]. Finally, depending on the shape of filter for the wavelet transform used in the decomposition during the encoding procedure, there is an exact downsampling relationship between the higher and the lower resolution shapes [8]. The relationship between corresponding object pixels at different resolutions should be maintained and considered as scalability constraint in the shape producing algorithms.

In this paper we present a semi automatic object extraction algorithm which is based on multiresolution spatial segmentation and backward region matching. The image at different resolutions is segmented with spatial scalability as a constraint. To extract enhanced shapes, border smoothness is also included in the objective function of spatial segmentation [9]. The objects of interest are determined by user intervention at the first frame and are tracked by a backward region matching which determines the objects' regions in subsequent frames. Regions with big matching error are further processed with a change detector to determine newly appeared objects/regions.

2. MULTIRESOLUTION VIDEO OBJECT EXTRACTION

2.1. Single resolution object extraction with down sampling

One regularly used option in video segmentation is the single level video segmentation where objects in fine resolution are extracted and then down sampled according to the existing relationship between shapes at different resolutions determined by the wavelet filter used [8]. However down sampling distort shapes and cannot preserve topology at lower resolutions for all possible shapes [10]. In other words, a visually pleasing object at higher resolution does not necessarily ensure similar quality at lower resolutions. For example in Fig 1, down sampling of two digital circles are compared.

It can be seen that better approximation of a digital circle at high resolution results in worse downsampled circle shape.

Extracting visually pleasing objects is an obvious objective in object extraction algorithms. Visual quality at high resolution degrades as we move towards lower resolutions; this is more pertinent for complex shapes with high perimeter to area ratio. In regular single resolution object extraction, followed by down sampling, this effect is not considered.

2.2. Scalable multiresolution object extraction with emphasis on smoothness

To produce more visually pleasing shapes we use smoothness, embedded in a multiresolution segmentation algorithm. It should be noted that the edge of most objects exhibits smoothness. Therefore, smoothness helps the extracted borders to resemble the normal objects/regions edges more closely. This will overcome the shortcomings of some region based segmentation algorithms in terms of border quality. To enhance the shapes at different resolutions, different coefficients are assigned to a smoothness function at different resolutions. A smoothness factor is defined by curvature estimation which depends on angles between adjacent border's pixels [11].

Smoothness enhances the shape especially in low contrast and jagged border areas. It deletes some pixels of foreground or reversely adds some pixels of background to objects that produces softer objects'/regions' borders which results in better shape view in all resolutions.

3. SPATIAL SEGMENTATION ALGORITHM

The proposed spatial segmentation fits MMRF segmentation with scalable object based wavelet coding. Images at different resolutions are segmented with spatial scalability as a constraint. By including border smoothness in the objective function of spatial segmentation enhanced shapes with better visual quality are extracted [9].

Different smoothness coefficients defined at different resolutions give some degree of freedom to put more emphasis on the low resolutions 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 an objective function matched with the problem at hand.

To extend the single level MRF based segmentation [12] to a multiresolution scalable segmentation algorithm, we note that the corresponding pixels at different resolutions have the same segmentation classification. Therefore the classification of these pixels changes together and they should be processed together in a multidimensional space. Consequently, objective function of regular single level Bayesian segmentation [12] is extended to a multidimensional space by the following equation:

$$E(X) = \sum_{\{S\}} \{\, ||Y(\{s\}) - \mu^{X(\{s\})}(\{s\})||^2 + \sum_{\{r\}\in\partial\{s\}} V_c(\{s\},\{r\}) + \sum_{q\in\{s\}} l_q \nu(q)\} \qquad (1)$$

In this expression, s is a pixel of the pyramid decomposition and $\{s\}$ is the set include s and its corresponding pixels on the other resolutions. Y, X and μ are intensity, segmentation classification and intensity average functions respectively. V_c is the clique function defined on two neighboring sets of corresponding pixels. In regular single level segmentation, cliques are defined over two adjacent pixels s and r by the following formula:

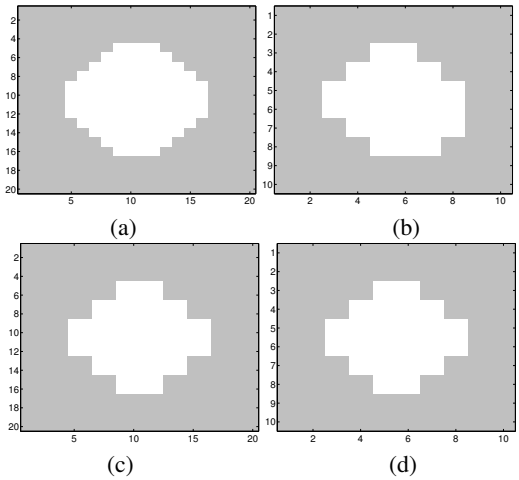


Fig. 1. (a) closer approximation of a digital circle at High resolution; (b) down sampling to lower resolution; (c) worse approximation of a digital circle at high resolution; (d) down sampling to lower resolution.

$$V_c(s,r) = \begin{cases} -\beta & \text{if} \quad X(s) = X(r) \\ +\beta & \text{if} \quad X(s) \neq X(r) \end{cases} \qquad (2)$$

But in the proposed scalable multiresolution segmentation algorithm, all the corresponding pixels of $\{s\}$ at different resolutions are tested with their neighboring pixels with the following equation:

$$V_c(\{s\},\{r\}) = \left(\frac{\sum_{k=M}^{M+N-1} L_K}{N}\right) \sum_{k=M}^{M+N-1} (-1)^{L_k}\beta\,,$$
$$L_k = \begin{cases} 1 \text{ if } X(s_k) = X(q_k) \\ 0 \text{ if } X(s_k) \neq X(q_k) \end{cases} \quad \& \;\; s_k \in \{s\},\; r_k \in \{r\} \qquad (3)$$

In (3), M is the lowest resolution in the pixels of $\{s\}$ and N is the number of different resolutions of pixels in $\{s\}$. A smoothness function at pixel q is shown by $\nu(q)$ where q is a pixel of set $\{s\}$. A smoothness coefficient is denoted by l_q which is resolution dependent. The first summation in (1) is over all pixel set of corresponding pixels at different resolutions and the second one is over the all cliques including the set $\{s\}$. The third summation is over all pixels of set $\{s\}$.

For the optimization of MMRF modelling, the Iterated Condition Mode (ICM) algorithm matched to the scalable multi resolution segmentation is used. The energy function of equation (1) is optimized sequentially from lower resolution to higher resolutions. In each resolution, in a raster scan order, the pixels are visited. At each pixel, by changing the segmentation classification for the processed pixel and its corresponding pixels, the energy function is optimized. After segmentation convergence at the current level the next higher resolution is processed. This repetitive algorithm continues until the finest resolution is achieved. This repetitive spatial segmentation algorithm can be found in [9].

The defined smoothness factor could be compared with the smoothness term in snake active contour model segmentation algorithms [11]. The main difference is that our approach is region

based while the active contour model is an edge based approach which has problems such as initial estimation and convergence to local optimum [11].

4. SEMANTIC VIDEO OBJECT EXTRACTION

At the core of most video segmentation algorithm routines is a tracking algorithm. In the backward tracking algorithm the spatial segmentation gives the precise borders of object(s). This also overcomes the problems of non rigid moving objects and uncovered background. Therefore we have proposed a multiresolution backward tracking algorithm.

In the first frame, through user's intervention and spatial segmentation, meaningful objects are determined. In the subsequent frames, the object is tracked by an automatic procedure. Multi resolution intra frame segmentation is performed as mentioned in Section 3. Scalable segmentation ensures similar segmentation patterns at different resolutions [9]. We have used this feature in our proposed tracking algorithm to track some regions in the proper resolution and extend the results to corresponding regions at other resolutions. Region classification starts from the lowest level of the decomposition. Regions bigger than a threshold are processed in this resolution and small size regions are processed in higher resolutions. Each processed region is divided into morphological catchment basins and each watershed basin is classified into object or background. Regions partitioning to basins overcome the probable short comings of spatial segmentation to separate the entire object from the background[1]. Motion estimation provides information for the backward projection of each basin. In this method, the translation motion vector is estimated for every processed region such as R of the current frame using the following formula which minimizes the matching error in RGB color space:

$$E = \min_{(u,v)} \sum_{(x,y)\in R} |I_t(x,y) - I_{t-1}(x+u, y+v)| \qquad (4)$$

In any projected region, if the number of pixels with the same label is more than a threshold, such as 50% of the projected region's size, it is classified the same as pixels' label. After processing the proper regions at the current resolution, the unclassified regions are processed in the higher resolutions. This repetitive algorithm continues until the finest resolution is achieved.

If the matching error of any processed region is larger than a threshold, region tracking is unsuccessful. In this case a modified change detector algorithm processes the region to detect the newly appeared objects/regions [3]. This algorithm, after a global motion compensation, computes the frame difference between two consecutive images. The idea of this approach is that the color/intensity variation of a moving region is different from that of stationary background because the motion of the moving object changes in color/intensity of the corresponding region [3]. Therefore, using a statistical test based on variance, we detect the regions with changed color or grey level. We denote the variance of color/intensity change for processed region by S_2 and the variance of the background color/intensity change by S_1. If the value of the following ratio is bigger than a threshold, a changed region is indicated.

$$V = \frac{S_2}{S_1} \qquad (5)$$

[1]depends on the example, 90% to $\%100$ of object area is correctly separated from background.

The variance S_1 is calculated on the background of the last frame.

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

5. EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate the performance of the presented algorithm, the three different sequences Clair, Hall monitor with CIF format, and Table Tennis with SIF format are segmented.

In the first example the tracking algorithm is run over the 75 frames of the Clair sequence. The extracted objects at frames number 20, 40 and 60 for different resolutions are shown in Figs. 2(a), (b) and (c). To compare the proposed algorithm with other region based object tracking and extraction methods, we have used similar tracking algorithm but in single resolution mode which includes regular single level spatial segmentation [12] and tracking only at the finest resolution. To ensure similarity to the existing region based tracking algorithms, which are often morphological based [1], the object areas were extended to fill the morphological catchment basins regions which overlap with the extracted object. The qualitative criterion for comparison is border smoothness of the extracted objects. Object smoothness is averaged over the curvature of border pixels. Although it is not an ideal criterion, it has confirmed performance of our subjective tests. The smoothness comparison for the 75 frames of the Claire sequence for the 3 resolution levels are shown in Table 1. The smoothness term affects the segmentation in areas of the image that have lower grey level contrast. In the Clair sequences the regions around the head have lower contrast compared to shoulder and body areas. If we only consider the head parts, the smoothness improves by 13.17%, 11.5% and 10.5% at different resolutions. As a qualitative example, Fig 3 shows the extracted objects of the $23th$ frame of the Clair sequence when using the scalable and a standard algorithm respectively. In this Fig, images of different resolutions are shown at the same size to highlight the details [2]. Analyzing both images, shows that our algorithm has extracted the Clair object smoother and more visually pleasing. It looks as if our algorithm has done a nice hair cut to Clair.

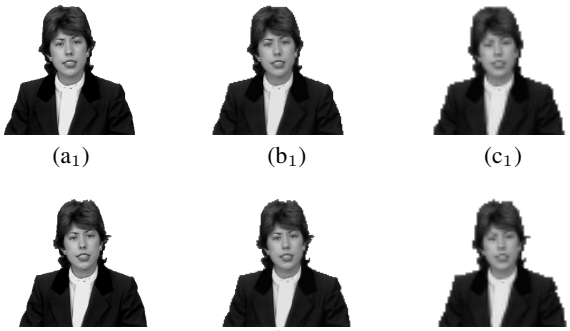
The proposed smoothed object extraction is different from a simple objects' border smoothness as has been done in [2] in the following areas. (1) Our smoothing process takes part in the segmentation algorithm and changes the segmentation outcome. (2) With sufficient contrast, the proposed algorithm produces borders that are more faithful to the object's shape. (3) On some occasions, some background pixels are added to the object to produce better looking shapes. (4) The smoothness factor could be adjusted for different resolutions to produce visually pleasing shapes at different resolutions.

As a second example, we have processed the standard MPEG-4 Table Tennis sequence, which has textured background with fast moving objects. Frame numbers 10, 20 and 32 with the extracted objects are shown in Fig 4. As an example, the extracted objects in frame number 10 of the table tennis sequence by the single level tracking expanded to watershed basins borders and the object extracted by our algorithm at 3 different resolutions are shown in

[2]Due to limited space, images' size are small and are not proper for a qualitative comparison. In the conference normal size and more examples will be shown.

(a) (b) (c)

Fig. 2. Claire sequence tracking; (a) Object extracted in frame number 20; (b) Object extracted in frame number 45; (c) Object extracted in frame number 65;



(a_1) (b_1) (c_1)

(a_2) (b_2) (c_2)

Fig. 3. Claire object of $23th$ frame; (a_1) scalable 288×352; (b_1) scalable 144×176; (c_1) scalable 72×88; (a_2) regular 288×352; (b_2) regular 144×176; (c_2) regular 72×88;

Table 1. Clair 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

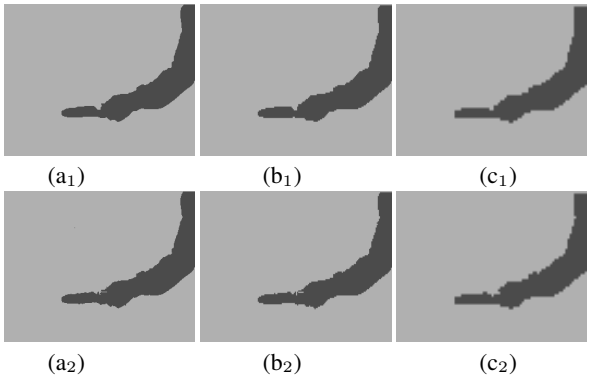
Fig. 5. For a quantitative comparison we have measured the object smoothness. The improvements are about 7% in different resolutions. Again if we only consider the hand and fingers with the racket, the smoothness improvements are nearly doubled. Also the time complexity of the multiresolution tracking algorithm is reduced to less than 30% of single resolution object tracking.

The proven high noise tolerance of the multiresolution image segmentation [9] is extended to video segmentation by the proposed algorithm. In video object extraction, moreover to it's spatial segmentation phase, noise can adversely affect the regions matching, especially at low contrast areas, resulting in wrong classifications. For example some small background regions close to the object area join the object and some regions belonging to object area join the background. To overcome this mismatching, we use lower resolutions to classify the regions. This effect is due to the wavelet transform used in pyramid decomposition filtering noise very well. Therefore, by decreasing the thresholds of regions' size, more regions are processed in noise-reduced environments at lower resolutions. The remaining unprocessed regions at high resolution are small regions which can be merged with the most similar neighbors.

To test the algorithm in noisy environments, a uniform noise

(a) (b) (c)

Fig. 4. Table Tennis object extraction; (a) frame 10; (b) frame 23; (c) frame 32;



(a_1) (b_1) (c_1)

(a_2) (b_2) (c_2)

Fig. 5. Table Tennis object $10th$ frame; (a_1) scalable 240×352; (b_1) scalable 120×176; (c_1) scalable 60×88; (a_2) regular 240×352; (b_2) regular 120×176; (c_2) regular 60×88;

(a) (b) (c)

Fig. 6. Object extraction from noisy Table tennis sequence; (a) frame $14th$ at resolution 240×352; (b) Scalable object extraction; (c) single level object extraction;

in the range $(-25, +25)$ is added to the tennis sequence. The noisy sequence is segmented with the proposed algorithm and the results are compared with single level tracking algorithm. The smoothness improvements are $10.8\%, 13.2\%$ and 14.6% respectively from low to high resolutions. The number of misclassified object's pixels for different resolutions for both scalable and regular video segmentation algorithms are counted in Table 2. The number of misclassified object's pixels in scalable multiresolution video segmentation algorithm decreases to 50% of pixels misclassification of regular single level segmentation algorithm. This confirms the superiority of the multi resolution algorithm. Fig. 6 shows the extracted objects in frame 18 for both multi resolution and single level object extraction.

In the third example, the Mall Monitor sequence is segmented. In this sequence object appears gradually and it cannot be determined by the user intervention at the first frame. Consequently, the change detector identifies new appeared objects/regions and the tracking algorithm detect the already appeared objects/regions. The extracted object of frame number 40 at different resolutions

Table 2. 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

using scalable segmentation algorithm are shown in Fig 7. In Fig 8, the extracted objects of frames number 34, 44 and 60 using the regular and the scalable algorithms can be seen. Some regions related to shade area are also detected as objects. Because the shades between two frames are also changed. Increasing change detector thresholds can reduce the detected shades but increases the risk of missing some parts of the object during the detection process. Fig. 8 confirms the superiority of the proposed algorithm over the regular object detection algorithm in creating a visually more pleasing segmentation. Table 3 confirms the improved smoothness of the proposed algorithm.

6. CONCLUSION

We have added a new quantitative criterion to region based video object extraction algorithms. This criterion qualitatively represents the visual quality of the objects at different resolutions. This criterion is a smoothness function based on the pixels' segmentation label and qualitatively represents the visual quality of the objects at different resolutions. To reduce the down sampling distortion, smoothness coefficients are considered for different resolutions. MMRF based spatial segmentation and tracking is used to extract the desired VOPs. The extracted objects are visually more pleasing and quantitatively smoother than objects detected through regular region based object extraction algorithms. The multiresolution algorithm has less computational complexity and can deal well with noisy environments. The produced shape's masks are directly usable for scalable wavelet based object coding.

REFERENCES

[1] Y.AverTsaig and A.buch, "Automatic segmentation of moving objects in video sequences: a region labeling approach," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 12, no. 7, pp. 597–612, 2002.

[2] F.Marques and J.Llach, "Tracking of generic objects for video object generation," in *International Conference on Image Processing (ICIP)*, 1998, vol. 3, pp. 628–632.

[3] Munchurl Kim, Jae Gark Choi, Daehee Kim, Hyung Lee, Myoung Ho Lee, Chieteuk Ahn, and Yo-Sung Ho, "A vop generation tool: automatic segmentation of moving objects in image sequences based on spatio-temporal information," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 9, no. 8, pp. 144–50, 1999.

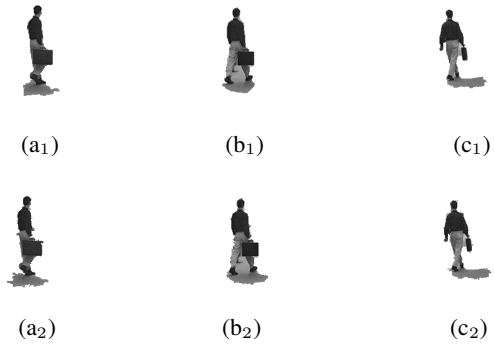
[4] ISO/IEC JTC1/SC29/EG11/N2322, "MPEG-4 Applications," Jul. 1998.

[5] De Freitas Mini, R. A.Campos, and M. F. M., "Visual tracking of objects using multiresolution," in *XII Brazilian Symposium on Computer Graphics and Image Processing (Cat. No.PR00481). IEEE Comput. Soc. 1999*, Los Alamitos, CA, USA, 1999, pp. 153–160.

[6] H. Danyali and A. Mertins, "Fully scalable texture coding of arbitrarily shaped video objects," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, 2003, pp. 393–6 vol.3.

(a) (b) (c)

Fig. 7. Hall monitor sequence Object extraction at frame $40th$; (a) resolution 288×352; (b) resolution 144×176; (c) resolution 72×86;



(a_1) (b_1) (c_1)

(a_2) (b_2) (c_2)

Fig. 8. Hall Monitor Sequence object extraction; (a_1) scalable extraction at frame 34; (b_1) scalable extraction at frame 44; (c_1) scalable extraction at frame 60; (a_2) regular extraction at frame 34; (b_2) regular extraction at frame 44; (c_2) regular extraction at frame 60;

Table 3. 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

[7] C. Christopoulos, A. SDkordas, and T. Ebrahimi, "The jpeg2000 still image coding system: an overview," *Consumer Electronics, IEEE Transactions on*, vol. 46, no. 4, pp. 1103–1127, Nov. 2000.

[8] A. Mertins and S. Singh, "Embedded wavelet coding of arbitrary shaped objects," in *Proc. SPIE 4076-VCIP'00*, Pert, Australia, June 2000, pp. 357–367.

[9] F. Akhlaghian Tab, G. Naghdy, and A. Mertins, "Multi resolution image segmentation with border smoothness for scalable object-based wavelet coding," in *Proc. 7th internatioinal conference on Digital Image Computing - Techniques and Applications (DICTA)*, Sydney, Australia, 2003, pp. 977–986.

[10] G.Borgefors, G.Ramella, G.Sanniti di Baja, and S.Svenson, "On the multiscale representation of 2d and 3d shapes," *Graphical Models and Image Processing*, vol. 61, no. 1, pp. 44–62, 1999.

[11] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1987.

[12] T. N. Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Trans. Image Processing*, vol. 40, no. 4, pp. 901–914, Apr. 1992.