Hybrid System to Moving Object Identification in Digital Video Clip

Dr. Tawfiq A. Abbas, Samaher Hussein Ali, Wafaa H. Al-Marsomi

Abstract— **Digital video has become very important form of information technology and is now used in many different areas, such as teleconferencing, mobile telephone, surveillance, and entertainment. Therefore, in this paper, we attempt propose a hybrid system to moving object identification in digital video clip. That system include four stages: the first stage is opening Audio/ Video Interleaved file (AVI) Video file which is called RIFF. AVI, second stage is segmented one of video frames using segmentation techniques presented in Thresholding technique. While the third stage used Discrete Wavelet Transform (DWT), The role of transform is to make the frame's energy as compact as possible, which means making just small number of transformed coefficients with large values while the magnitude of most of the rest coefficient is quite small, or zero. It should be noted that transform itself does not produce any compression, instead, its task is to produce a format, then extraction features from wavelet transform Coefficients includes (Momentum Features) and we used these features as inputs to training neural network. The fourth stage which is used to identify the moving object in digital video clip using Radial Basis Function (RBF) Network.**

***Index Terms*— Audio/ Video Interleaved file, Wavelet Transform, Object Isolation, Radial Basis Function Network.**

I. Introduction

MANY evolving multimedia applications require transmission of high quality video over the network. One obvious way to accommodate this demand is to increase the bandwidth available to all users. Of course, this "solution" is not without technological and economical difficulties. Another way is to reduce the volume of the data that must be transmitted. There has been a tremendous amount of progress in the field of video compression during the past 10 years. In order to make further progress in video coding, many research groups have begun to use wavelet transforms. In this paper, we will briefly discuss the nature of wavelets, and some of the salient features of image and how we can identify objects of digital video using neural networks. Since this is a very rapidly evolving field, only the basic elements are presented

Tawfiq A. Abbas is with the Computer Science Department , University of Babylon, Iraq ; (e-mail: tawfiqasadi63@yahoo.com).

Samaher H. Ali is with the Computer Science Department, University of Babylon, Iraq ; (e-mail: samaher_hussein@yahoo.com).

Wafaa H. AL-Marsomi is with the Law College, University of Karbala, Iraq; (e-mail: wafsam2005 @yahoo.com).

II. Object Isolation (Segmentation)

Segmentation is the first step in the frame analysis. It refers to subdividing a frame into distinct regions that are supposed to correlate strongly with objects or features of interest in the frame.

In general, two conditions should be fulfilled in image segmentation [1] :-

- Pixels that are grouped together (that belong to the same region) must have similar attributes.
- Generated regions (groups of pixels) should be meaningful.

Let I denote a frame (image), and let P be a logical predicate which is defined as partition S=R1, R2, …, Rn of I that verifies the following conditions[2]:

- $\bigcup_{i=1}^{N} R_i = I$
- R_i is connected $\quad i=1,2,\ldots n$
- $P(R_i)=\text{True} \quad i=1,2,\ldots n$
- $P(R_i \cup R_j)=\text{False} \quad i \neq j$ for all adjacent regions R_i , R_j.

In this definition, the first condition implies that segmentation is complete (i.e., every pixel should belong to a region).

The second condition requires that pixels in a region are connected (i.e., regions are composed of contiguous pixels).

The third condition determines what kind of properties the segmented regions should have (i.e., confirms that every region is homogenous and verifies the similarity criteria).

The fourth condition expresses the maximality of each region in the segmentation (i.e., affirms that a region could not be extended any more).

Verifying these conditions is considering a very difficult mission. Because real world frames contain some complexity due to overlapping objects and high contrast amissibility between these regions.

In general, the segmentation process has two types of mistakes:

• The segmentation process has added new regions (i.e., region not represented actual objects in image).

• Some regions may be amissbile in region.

Therefore a fixed theory is not found in relation to segmentation problem, where all techniques found depend on ad-hoc principles in performance to segmentation process.

The following are the most popular methods in segmentation process [1] [2]: Thresholding, Boundary Detection, Region growing, Region splitting and Merging, Clustering Techniques. in this work, we deal with the threshold techniques.

III. Wavelet Transform

Most of the signals in practice are *time-domain* signals. When we plot time-domain signals, we obtain a time-amplitude representation of the signal. In many cases, the most important information is hidden in the frequency content of the signal. The frequency spectrum of a signal shows what frequencies exist in the signal.

Wavelet transform provides better time-frequency localization than other trans-form tools. In literature, many paper have proven that it is extremely efficient for transform-based image compression. From our or other researchers experimental work, it shows that many of wavelet transform coefficients for a typical image tend to be very small or zero, making those coefficients more easily to be coded. In [3], the author pointed out the underling reason about why wavelet is so powerful for image compression. It is believed that the basic functions associated with a wavelet decomposition typically have both long and short support. The basic functions with long support are effective for representing slow variations in an image while basic functions with short support can efficiently represent sharp transitions (i.e., edges). This makes wavelets ideal for representing signals having mostly low- frequency content mixed with a relatively small number of sharp transition. With more traditional transforms techniques like DFT and DCT, the basis functions have support over the entire image, making it difficulty to represent both slow variations and edges efficiently. By using wavelet transform recursively, the whole frame is divided into several parts, with each part represents different bandwidth version of original frame. a logarithmic advantage is that the low-frequency bands have small bandwidths, while high-frequency bands have large bandwidths. This phenomena matches human visual perception behaves. So that, the information that have large effects of human visual perception tend to be compact, making it easy to code.

The continuous and discrete wavelet transforms are given in (1) and (2), respectively [4]

$$(T^{wav} f)(a,b) = |a|^{-1/2} \int f(t)\psi(\frac{t-b}{a})dx \qquad (1)$$

$$T^{wav}_{m,n}(f) = a_0^{-m/2} \int f(t)\psi(a_0^{-m}t - nb_0)dt \qquad (2)$$

in this paper, we used the discrete wavelet transform as explain in the stages of proposed method.

IV. Radial Basis Function Network

Radial Basis Function emerged as a variant of artificial neural network in late 80 's. However, their roots are entrenched in much older pattern recognition techniques as for example potential functions, clustering, functional approximation, spline interpolation and mixture models. RBF's are embedded in a two layer neural network, where each hidden unit implements a radial activated function. The output units implement a weighted sum of hidden unit outputs. The input into an RBF network is nonlinear while the output is linear.

In order to use a RBF network we need to specify the hidden unit activation function, the number of processing units, a criterion for modeling a given task and a training algorithm for finding the parameters of the network. Finding the RBF weights is called network training.

RBF networks have been successfully applied to a large diversity of applications [5] including interpolation, electronic device parameter modeling, channel equalization, speech recognition, image restoration, 3-D object modeling, motion estimation and moving object segmentation, data fusion, etc. Various function have been tested as activation function for RBF network while in this paper, the Gaussian function is preferred. Mixtures of Gaussian have been

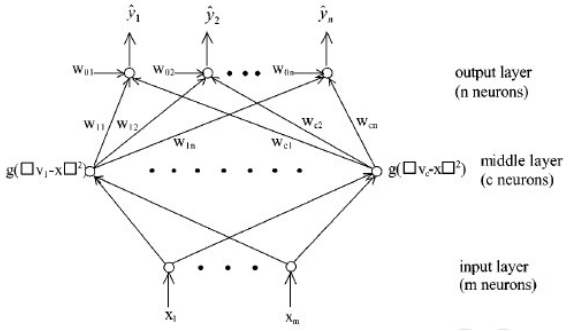

Fig. 1. Radial Basis Function Network Architecture.

considered in various scientific fields. The Gaussian activation function for RBF network is given by:-

$$\phi_j(X) = \exp\left[-(X-u_j)^T \sum_j^{-1} (X-u_j)\right] \qquad (3)$$

For j=1,…, L, where X is the input features vector, L is the number of hidden units, uj, Σ_j are the mean and the covariance matrix of the jth Gaussian function. In certain approaches a polynomial term is added to the expression (3), while others the Gaussian function is normalized to the sum of all the Gaussian components as in the Gaussian- mixtures estimation.

The output layer implements a weighted sum of hidden-unit outputs:

$$\psi_K(X) = \sum_{j=1}^{L} \lambda_{jk} \phi_j(X) \qquad (4)$$

For k=1,…,M where λ_{jk} are the output weights, each corresponding to the connection between a hidden unit and an output unit and M represent the number of output units. The weight λ_{jk} show the contribution of a hidden unit to the respective output unit. In a classification and identification problems if λ_{jk} >0 the activation field of the hidden unit j is contained in the activation field of the output unit k.

In the pattern identification applications, the output of the radial basis function is limited to the interval (0,1) by sigmoid function:

$$Y_k(X) = \frac{1}{1+\exp\left[-\psi_k(X)\right]} \qquad (5)$$

IV. The Proposed Method

Moving Object Identification in Digital Video Clip is one of the important areas in computer vision. This part of paper assignees to represent the structure of the proposed method and show how we can implementing this method to identify any object in digital video frame. Figure 2 explains the proposed method.

A. Open AVI file

In this stage we used algorithm to open AVI structure and search to list movie (i.e. sequence frames/images) which are needed in the next stages to recognition it. In this algorithm the
function (ASCII to long)was used to convert FCC ASCCII character to long for example convert FCC RIFF to long by calling ASCII to long (“R”,”I”,”F”,”F”) to open AVI file, then after that take information from AVI header such as total frame, height, and width of the image and then search for list movie using search function about key word “MOVI”, when we search make a loop from one to sequence of image(frames) then search about keyword “00db” which is the initial of each frame, then make two loops from height and from width respectively to read RGB pixel image, and split it into three bands that save it into three individual matrices called (red, green, blue) which are used in the next stage[6] .

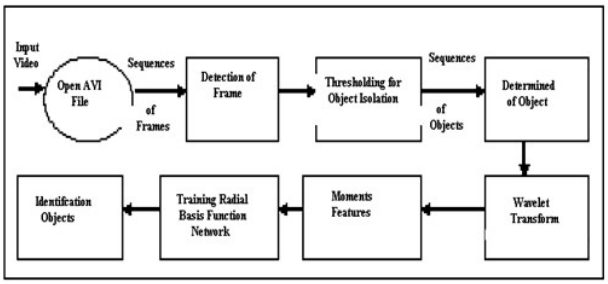

Fig. 2. A Block Diagram of The Proposed Method

B. Detection of frame

In this stage, we determine the frame that our wish to work on it from the sequence frames.

C. Thresholding for Object Isolation

The Thresholding technique is very popular in image processing operations, one of which is the segmentation (object isolation). Thresholding transforms a dataset containing values that vary over some range into a new dataset containing values that vary over a smaller range. The simplest case when the destination dataset contains only two values; a threshold is applied to the input data so that values falling below the threshold are replaced by one of the values in the output dataset; input values at or above the threshold are replaced by the other output value.

Threshold techniques in image processing are based on the threshold values, which are usually selected from the image histogram. This operation encounters no problem if either the portion of the image area occupied by the objects is known or the gray level ranges of objects and background are well separated (the gray level histogram has deep valley between two peaks)[2].

In general, a single threshold value is not enough to detect all the objects in a complicated frame. Therefore, it is more efficient to use multiple thresholds rather than a single threshold.

Multiple Thresholding is an operation that involves tests against a D-dimensional function T

$$(T_0,T_1,\ldots,T_{D-1}) = T\{X_R,Y_R,P(X_R,Y_R),F(X_R,Y_R)\}$$

when, we work on red matrix of frame.

$$(T_0,T_1,\ldots,T_{D-1}) = T\{X_G,Y_G,P(X_G,Y_G),F(X_G,Y_G)\}$$

when, we work on green matrix of frame.

$$(T_0,T_1,\ldots,T_{D-1}) = T\{X_B,Y_B,P(X_B,Y_B),F(X_B,Y_B)\}$$

when, we work on blue matrix of frame.

The resulting threshold frame I(X,Y) use D+1 values (V_0,V_1, …,V_D) to map D+1 objects in the frame defined by the D thresholds:

$$I(X, Y) = \begin{cases} V_0 & IF \quad F(X,Y) <= T_0 \\ V_1 & IF \quad T_0 < F(X,Y) <= T_1 \\ V_D & IF \quad F(X,Y) > T_{D-1} \end{cases} \quad (6)$$

D. Determined of segment (object)

In this stage of the proposed method, we determined which object of the sequences of object need to known moving in all the other frame. This stage is very important in proposed method.

E. Discrete Wavelet Transform

Wavelets are generated from on single function f by dilations and translations[7].

$$\varphi_{s,u}(x) = 2^{s/2}\varphi(2^s x - u) \quad (7)$$

Using the wavelet transform, any signal f can be represented as a superposition of wavelets.

$$f(x) = \sum_{s,u \in Z} W_{s,u}\varphi_{s,u}(x)$$
$$W_{s,u} = \int \varphi_{s,u}(x) f(x) dx \quad (8)$$

$W_{s,u}$ is the wavelet representation of the signal.

Wavelet transform provides a hierarchical signal representation, each coefficient corresponds to a spatial area and a frequency range. It is identical to a hierarchical subband system, where the subbands are logarithmically spaced in frequency and represent octave-band decomposition. At each level, the signal can be further decomposed into a coarser approximation and a corresponding added detail Figure 3 explains that.

The coefficients in the same orientation and corresponding to the same spatial region in the image can be organized as a tree, where each parent node has four children nodes, which are in the higher frequency band corresponding to the same spatial region. Such tree-structured representation provides and efficient means for exploitation of wavelet coefficients clustered both in spatial and frequency domain.

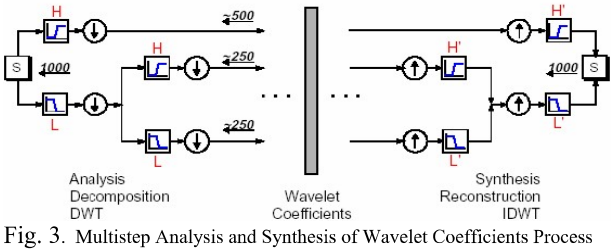

Fig. 3. Multistep Analysis and Synthesis of Wavelet Coefficients Process

Fig. 4. A One-Level Wavelet Decomposition

At the first stage of wavelet transform, the original frame is divided into four subbands by separable application of vertical and horizontal filters. Each coefficient represents a spatial area corresponding to approximately a 2×2 area of the frame. The low frequencies (L) represent $0<|?|<p/2$, and the high frequencies (H) represent $p/2<|?|<p$, as shown in Figure 4. The HL1, LH1 and HH1 represent the finest scale wavelet coefficients.

Then, the subband LL1 is further decomposed by separable vertical and horizontal filters, as shown in Figure 5. This process is continued till the final level is reached. For example, a 256×256 frame will have 8 levels of subbands. At each level, there are 3 subbands, and a remaining low frequency subband representing all coarser scales. The coefficients in higher level, or say coarser scale, represent larger spatial area but a narrower frequency band. In subband coding systems, the coefficients from a given subband are usually grouped together for the purposes of designing quantizers and coders.

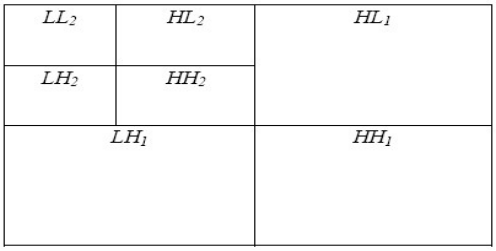

Fig. 5. A Two-level Wavelet Decomposition

F. Moments Features

The task of object identification that not depend on the change which happened on the object as the change in the size and position is impact on the objects features. So the concern appear to using moments features and specially the central moments because it has invariant features with the changes that happened to on the frame as translation, rotation ,scaling and deflection The extraction moments for frame or objects which has size N×M and q,p values which represent the moment rank and its values (q,p=0,1,2,3,4…..n), F(x,y) represent the intensity of object in coordinated x,y where x=0,1,2,…,N-1, Y=0,1,2,…,M-1[8].

We can compute the ancentral moment as follow:

$$M_{pq} = \sum_{X=0}^{N-1}\sum_{Y=0}^{M-1} X^P Y^q F(X,Y) \quad (9)$$

While the central moment compute as follow:

$$\mu_{pq} = \sum_{X=0}^{N-1}\sum_{Y=0}^{M-1}(X-\overline{X})^p(Y-\overline{Y})^q F(X,Y) \quad (10)$$

Where

$$\overline{X} = \frac{M_{10}}{M_{00}},\ \overline{Y} = \frac{M_{01}}{M_{00}}$$

In this paper, we deal with p=0,1,2. and q=0,1,2,3. therefore the number of features extraction from objects equal 12 feature. And theses features extraction from the sub band LL1 results from wavelet transform. in this work, we ignore the details coefficients (HH,HL,LH) and enough by the features get from sub band LL1.

*G. Training **Radial Basis Function Network***

By means of training, the neural network models the underlying function of a certain mapping. In order to model such a mapping we have to find the network weights and topology. There are two categories of training algorithms: supervised and unsupervised. RBF networks are used mainly in supervised applications. In a supervised application, we are provided with a set of data samples called training set for which the corresponding network outputs are known. In this case the network parameters are found such that they minimize a cost function:

$$\min \sum_{i=1}^{Q} Y_K(x_i) - F_k(x_i))^T (Y_k(x_i - F_k(x_i)) \quad (11)$$

where Q is the total number of vectors from the training set, $Y_k(x_i)$ denotes the RBF output vector and $F_k(x_i)$ represents the output vector associated with the a data sample X_i from the training set.

The centers of the radial basis function are initialized randomly. And these centers are update as follows:

$$\hat{\mu_j} = \hat{\mu_j} + \eta(x_i - \hat{\mu_j}) \quad (12)$$

where η is the training rate. For a minimal output variance, the training rate is equal to the inverse of the total number of data samples associated to that hidden unit.

V. RESULT

In System of moving object Identification in digital video clip three video file movies were taken which are different in natural complexity, scenes quality, and size.

By applying the proposed method on these movie we noticed the identification ratio still highly and changeless although different complexity of video file movies this point explain the accuracy of the proposed method as explain in the following case studies :

A. Case Study Number One

Movie number one represents personal cross river in the forest and which contain 102 frames distributed over several scenes such as tree, river, road, personal, sky, grass,…,ect. We applied the Thresholding technique to segmentation each frame into the objects then determined the personal as object (to test the proposed method and explains how we can known (identify) these object in different frames although change in there position. After that, we used the discrete wavelet transform to reduce the domain of search depended on the equations represented in paragraph (E from IV). Then find the 12 features from the approximation coefficient of the discrete wavelet transform by using the momentum equations as explain in equation (9, 10). The final step, training RBF network on these features but before this. We need to determined the topology of the network and sum of the parameters: in this case study, number of frame used in training equal 32, number of unit in input layer equal 12, number of unit in hidden layer 8, learning factors equal 0.007. while the number of frames used in test stage equal 70.

B. Case Study Number Two

Movie number two represents players in Football Stadium and which contain 305 frames distributed over several scenes such as players, ball, Scrolls, grass, encouragers…,ect. We applied the Thresholding technique to segmentation each frame into the objects then determined one player as object (to test the proposed method and explains how we can known (identify) these object in different frames although change in there position. After that, we used the discrete wavelet transform to reduce the domain of search depended on the equations represented in paragraph (E from IV). Then find the 12 features from the approximation coefficient of the discrete wavelet transform by using the momentum equations as explain in equation (9, 10). The final step, training RBF network on these features but before this. We need to determined the topology of the network and sum of the parameters: in this case study number of frame used in training equal 55 , number of unit in input layer equal 12, number of unit in hidden layer 25, learning factors equal 0.005. while the number of frames used in test stage equal 250 frame.

C. Case Study Number Three

Movie number three represents monkey jumps and plays in enguarded and which contain 94 frames distributed over several scenes such as monkey, trees, grass, bananas…,ect. We applied the Thresholding technique to segmentation each frame into the objects then determined a monkey as object (to test the proposed method and explains how we can known (identify) these object in different frames although change in there position. After that, we used the discrete wavelet transform to reduce the domain of search depended on the equations represented in paragraph (E from IV). Then find the 12 features from the approximation coefficient of the discrete wavelet transform by using the momentum equations as explain in equation (9, 10). The final step, training RBF network on these features but before this. We need to determined the topology of the network and sum of the parameters: in this case study number of frame used in training equal 20 , number of unit in input layer equal 12, number of unit in hidden layer 9, learning factors equal 0.01. while the number of frames used in test stage equal 74 frame.

VI. Conclusion

In this study we provide an introduction to hybrid system moving object Identification in digital video clip have very attractive properties such as accuracy, generality, functional approximation, interpolation. Theses properties made them attractive in many applications. Very different fields (not only in identification the moving objects of digital video clip) such as signal and image processing, computer vision used them successfully for various tasks. We present some examples when applying hybrid system to identification different objects including personal cross river in the forest, player moves in Football Stadium and monkey plays and jumps in enguarded. In all the above cases studies the proposed system is successful in there task and given highly accuracy.

References

[1] Scott.E, "Computer Vision and Image Processing ":A Practical Approach Using CVIP Tools", Prentice-Hall ,New Jersey ,1998.

[2] Gonzalez .R and Woods. R," Digital Image Processing (2nd ed.)", New Jersey: Prentice Hall, 2002.

[3] D. L. Williamson, J. B. Drake, J. J. Hack, R. Jakob, and Swarztrauber P. N. A stan-dard test set for numerical approximations to the shallow water equations in spherical geometry. J. Comput. Phys., 102:211{224, 1992.

[4] Yin Xiaowei, "Wavelet Techniques for Color Document Image Coding," PD.H thesis, Department of Electronic System Engineering University of Essex, 2004.

[5] Wafaa. H.A. "Video clip compression using DCT technique" Dep. Of computer science, University of Babylon, 2006.

[6] H. Sun and Yi Zhang, "Embedded Zerotree Wavelet Image Codec", University of Wisconsin – Madison Electrical Computer Engineering, 2003

[7] Adrian G. Bors, " Introduction of the Radial Basis Function Networks", Determent of computer Science, University of York, York, YO105DD, UK, 2001.

[8] Jain Kl. K, "Fundamental of Digital Image Processing", Prentice Hall, India, 2000.