# Perception-Oriented Prominent Region Detection in Video Sequences

Lang Congyan, Xu De and Yang Xu
School of Computer Science & Information Technology,
Beijing Jiaotong University,
Beijing, 100044, China

E-mail: gltree@263.net, xd@comput.njtu.edu.cn

**Keywords:** prominent region, image segmentation, feature extraction, video content representation

**Received:** October 21, 2004

*Effective and efficient representation of video sequences is an important yet challenging task for video retrieval and browsing. In this paper, we propose a new approach for the prominent region detection from the viewpoint of the human perception intending to construct a good pattern for content representation of the video sequences. Firstly, we partition each frame into homogeneous regions using a technique based on a non-parameter clustering algorithm. Then, in order to automatically determine the prominent importance of the different homogenous regions in a frame, we extract a number of different mise-en-scene-based perceptual features which influence human visual attention. Finally, a modified Fuzzy Inference Neural Networks is used to detect prominent regions in video sequence due to its simple structure and superior performance for automatic fuzzy rules extraction. The extracted prominent regions could be used as a good pattern to bridge semantic gap between low-level features and semantic understanding. Experimental results show the excellent performance of the approach.*

*Povzetek: Predstavljen je nov postopek za zaznavanje regij.*

## 1 Introduction

With the current advance of video database technique, efficient video retrieval and browsing have become crucially important, especially with the development of the video content description standard, such as MPEG-7. Unfortunately, current approaches to video processing suffer from one or more shortcomings that stem from the semantic gap between low-level features and high-level semantic concepts.

To bridge the semantic gap, most previous works select semantic video objects [1-3] as the underlying video patterns for video content representation and feature extraction. However, the major problem using semantic video object as video patterns is that automatic semantic video object extraction in general still needs for human's interaction at the current stage.

Faced with these problems, an increasing number of researchers are now exploring the intermediate-level processing, shifting the focus of attention away from the purely local and pixel-based indicators to more global measures that seem to provide a stepping stone towards a more robust high-level processing [18]. Many image and video processing applications could be made both more efficient and effective if a number of salient regions were first segmented.

Studies of visual attention and eye movements [4,5] have show that humans generally only attend to a few areas in an image. Even when given unlimited viewing time, subjects will continue to focus on these few areas rather than scan the whole image. According to the fact, many research efforts have been given in detecting salient region in image intending to overcome the limitations of semantic object extraction. A considerable amount of research has addressed the salient region detection problem by clustering-based methods, for instance, in Ref [18], authors firstly map an image into the appropriate feature space, then detection salient regions by nonparametric clustering method. Hang Fai Lau, et al [19] identify a small number of regions in an image using low-level features, which work well on the colour image for image retrieval. On the other hand, most existing approaches [6,7] aim at detecting the salient region for images, which are mainly based on the construction of a saliency map modeled as an integration of different measurable, low-level image features such as color, orientation, depth information etc. The purpose of the saliency map is to represent the conspicuity of each locations of the visual field, that is, salient regions extracted have higher prominent importance than the other regions. In [11], authors use motion information to construct salient map for video sequence, which gives superior performance for moving region analysis in video sequence. A salient region detection and tracking method is presented in [15], which extract salient regions based on color and orientation maps followed by a tracking mechanism.

Salient region extraction based on saliency map provides a good starting point for semantic-sensitive content representation. However, perceived salient region extraction for image or video is still an unsolved problem. One reason is that video sequence has more

context information than single image, hence, low-level features are often not enough to classify some regions unambiguously without the incorporation of high-level and human perceptual information into the classification process. Another reason for the problems is perception subjectivity. Different people can differ in their perception of high-level concepts, thus a closely related problem is that the uncertainty or ambiguity of classification in some regions cannot be resolved completely based on measurements methods. A basic difference between perceptions and measurements is that, in general, measurements are crisp whereas perceptions are fuzzy [17].

In this paper, we propose a new method for prominent region extraction in video sequences in order to remove limitations explained above. For each frame, a pre-segmentation composed of homogeneous regions is produced, and then the segmented image is analyzed by a number of perceptual attributes based on the *mise-en-scene* principles. As a set of techniques, *mise-en-scene* helps compose the film shot in space and time [8], which are used by the filmmakers to guide our attention across the screen, shaping our sense of the space that is represented and emphasizing certain parts of it .

It is known that fuzzy logic can provide a flexible and vague mapping from linguistic terms to represent the human perception, and neural networks have superior learning ability to extract fuzzy rules automatically. Hence, to enable alleviate the semantic gap and the perception subjectivity problems, our method for automatically determining the perceptual importance of regions is constructed based on fuzzy inference neural networks(FINNs).

While most existing work focus on the detection of salient region, our approach for extraction of perception prominent regions is distinctive with several important advantages: (1) According to the *mise-en-scene* principles, the perceptual features are extracted for homogenous regions, rather than the low-level features, so as to provide more encouraging pattern to classifier; (2) The prominent importance of regions is assigned through soft decisions. Experiments show the effectiveness and robustness of our method on different type of video.

The rest of the paper is organized as follows: Pre-segmentation process of image frames is described in the next section. In Sect. 3, the perceptual feature extraction for primitive homogenous regions is implemented. And then, prominent region detection based on FINNs is presented in Sect.4. The effectiveness of the proposed approach is validated by experiments over real-word video clips are expressed in Sect.5. Concluding remarks are given in Sect. 6.

## 2 Pre-segmentation of Image Frames

As stated in Ref [19], non-parametric density estimation techniques are well suited to the task of segmenting coherent regions in an image. Therefore, each frame is initially segmented into homogeneous regions based on mean shift algorithm, the color and texture information is incorporated into the mean shift segmenter [9] in this section. To segment an image, we first partition it into 4*4 blocks and extract a 6-D feature vector for each block. In particular, three of them are the average color components computed in CIE LUV color space due to its perceptually uniform derivation of the standard CIE XYZ space. The other three represent energy in high frequency bands of wavelet transforms [14]. And then, the 3 wavelet features are computed as the square root of the $2^{nd}$-order moment of the wavelet coefficients in the HL, LH, and HH frequency bands.

The wavelet image decomposition provides a representation that is easy to interpret. Every subimage contains information of a specific scale and orientation, spatial information is retained within the subimages and the coefficients in different frequency bands show variations in different directions. We use the Daubechies discrete wavelet transform to decompose the image data into wavelet coefficients. After extract the color and texture features, they must be combined to form a single feature vector. Concerned the dynamic range of each feature and its relative importance, all features must be normalized and weighted. Thus, an integrated feature vector is formed as follows:

$$f_{ct-block}(i) = (w_{color}V_c, w_{texture}V_T) \quad (1)$$

$$V_c = \{C_1, C_2, C_3\}; V_T = \{T_1, T_2, T_3\};$$

where $w_{color}$ *and* $w_{texture}$ are the weights for color and texture selected experientially. $V_c$ *and* $V_t$ are the extracted features color and texture, respectively.

After the 6_D feature vector is extracted for all 4*4 blocks**,** we apply the mean shift clustering approach [9] to segment the image into homogenous regions.

Fig.1 shows the results of partitioned homogenous regions for two frames. The level of accuracy in this initial step is important for the overall performance of the proposed prominent region detection, as these pre-segmented homogenous regions will constitute the basic

(a) (b)

**Fig.1.** Homogenous region segmentation examples: (a) original images; (b) respective homogenous region

contours of the extracted perceptual prominent regions. For that reason, the parameters of this initial step are selected so that the frames are over-segmented rather than under-segmented.

# 3 Region Feature Description

Human selective attention makes us distinguish important features of input stimulus from others and focus on a critical feature of an input stimulus. Most basically, our visual system is attuned to perceiving change both in time and space, since one of the main goals of human visual system is to minimize uncertainty [5]. This is also in agreement with Gestalt Organization Theories. By taking advantage of these facts, the filmmaker uses the arrangement of the *mise-en-scene* to attract our attention by means of changes in light, shape, movement and other aspects of the image [8]. Thus, in order to get suitable content pattern of region in video sequences, we extract a number of perceptual features described below.

### *1) Contrast of region with surroundings (CSr)*

Regions, which have a high contrast with their surroundings, are likely to be greater visual importance and attract more attention. The filmmaker can exploit principles of color contrast to shape our sense of screen space. For instance, bright colors set against a more subdued background are likely to draw the eye [8]. The contrast importance $CSr(R_i)$ of a region $R_i$ is calculated as:

$$CSr(R_i) = \sum_{m=1}^{n} I^*(R_i) - I^*(R_{i-neighbours_m}) \quad (2)$$

where $I^*(R_i)$ is the mean intensity of region $R_i$, and $I^*(R_{i-neighbours_m})$ is the mean intensity of the *m-th* neighboring regions of $R_i$.

### *2) Orientation Conspicuity of Region (OCr)*

Gabor filtering allows to get information about local orientation in the image, thus orientation map computed by Gabor filtering is an important recognition cue, which was chosen on the basis of the evidence from the study of human attention [10]. Here, this technique is also employed to descript region orientional information importance.

Local orientations $O_\theta$ are obtained by applying Gabor filters to the original images from particular orientations $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$. The four oriented images are added up to generate the desired orientation map, and then normalized orientation map $\overline{I}_{orientations}$ is achieved by using a traditional normalization operator.

$$OCr(R_i) = \frac{\sum_p \overline{I}_p}{N_{pixel}(R_i)}, p \in R_i \quad (3)$$

where $N_{pixel}(R_i)$ denotes the number of pixels in the region $R_i$.

### *3) Shape Indicator of Region (SIr)*

The shape indicator of region can be calculated as:

$$SIr(R_i) = \frac{N_{edge}(R_i)}{N_{pixel}(R_i)} \quad (4)$$

where $N_{edge}(R_i)$ is the number of pixels in the region $R_i$ which border with other regions. Generally speaking, a small $SIr(R_i)$ signifies a long, thin region, which further implies a high prominent importance, while for rounder regions it will be lower.

### *4) Compositional Balance Indicator of Region (CIr)*

Compositional Balance is an importance factor of the *mise-en-scene* to guide our attention and shape our viewing of the image, it can be interpreted as the extent to which the areas of screen space have equally distributed masses and points of interest. Generally the filmmakers often center the frame on the character's body which is a simplest way to achieve compositional balance [8]. However, no existing works pay attention to this information. Based on the above observations, the compositional balance indicator is determined according to the following measure:

$$CIr(R_i) = \begin{cases} \dfrac{CSr(R_i)}{\left\| gc(R_i) - \overline{gc(R)} \right\|}, \overline{gc(R)} \in R_i \\ \dfrac{CSr(R_i)}{\left\| CSr(R_i) - CSr(R_i') \right\| + \left\| gc(R_i) - \overline{gc(R)} \right\|}, \overline{gc(R)} \notin R_i \end{cases} \quad (5)$$

where $gc(R_i)$ is the gravitational center of the region $R_i$ and the mean of gravitational center for all regions in image is denoted as $\overline{gc(R)}$. And the region $R_i'$ is selected whose gravitational center is the nearest neighbor of the symmetrical point of $gc(R_i)$ with respect to the midline of the frame. If the character's body is located in the frame center, we know that the

larger *CSr* and the nearer distance between its gravitational center and $\overline{gc(R)}$ the region in image is, the larger *CIr* the region is, meaning that the higher possibility that it will be a character portion of the frame. For the second case, as the same sense, the higher *CIr* shows that the frame may balance two or more elements encouraging our eye move between these regions.

#### *5) Motion Prominent Indicator (MIr) of region*

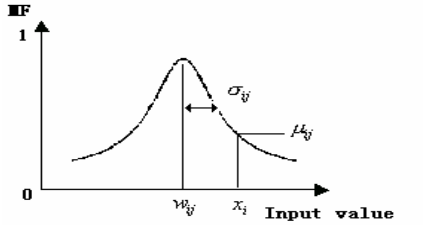
Almost invariably, a moving item draws our attention more quickly than a static item does. Motion information is an important cue for human to perceive video content. In our work, block motion vectors are used for motion analysis.

Since of a motion vector reflects motion velocity of the block, we define the $IM(R_i)$ as the mean of the magnitude of the motion vectors in the region $R_i$. An approach similar to [11] is adopted to descript our motion consistency of region, denoted as $MC(R_i)$. Specially, the motion consistency is calculated based on the entropy, which should be consulted for the details in [11], estimation of the probability is obtained by summation over rows and columns of eight-bin phase histogram of the region $R_i$. Then we define the motion prominent importance as following:

$$MIr(R_i) = IM \square MC \qquad (6)$$

# 4 Prominent Region Detection based on FINNs

After the region features are extracted, the perceptual prominence is required to assign to each region. Fuzzy logic has the capability of modeling perception vagueness, uncertainty and can support human-like reasoning. On the other hand, studies of fuzzy neural networks that combine both advantages of the fuzzy systems and the learning ability of the neural networks have been carried out. These techniques can alleviate the matter of fuzzy modeling by learning ability of neural networks [12, 13]. FINNs is ideal for our problem as it can extract and use rules automatically with a simple structure and superior performance.



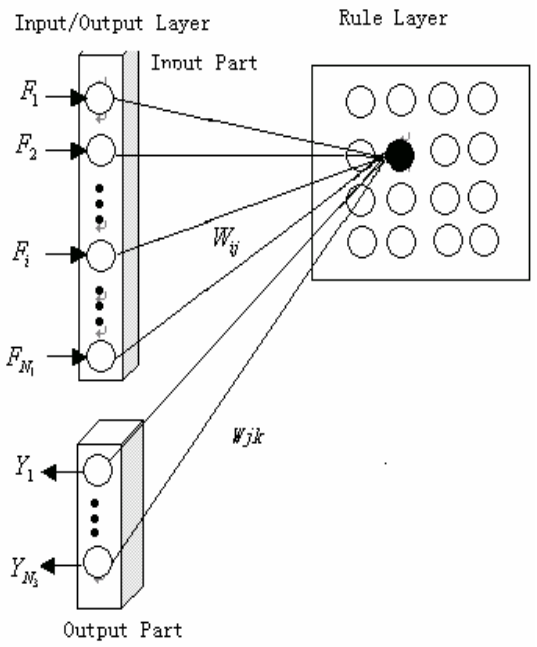




**Fig.2** the structure of FINNs and its membership function.

## 4.1 Architecture of FINNs

Fig.2 shows the basic structure of FINNs [12]. It consists of two layers. One is the input-output (I/O) layer and another is the rule-layer. The I/O layer consists of the input- and the output- part. Each node in the rule-layer represents one fuzzy rule. Weights from the input-part to the rule-layer and those from the rule-layer to the output-part are fully connected and they store fuzzy if-then rules. The number of neurons in the input-part is equal to the dimension $N_1$ of the input data, the number of rules is $N_2$, and $N_3$ is the number of output node**.**

For the prominent region detection problem each segmented region was described using a set of five dimension features, comprising of the perceptual features defined in section 3. Let $F = (F_1, F_2, F_3, F_4, F_5)$ denote the perceptual features *CSr*, *OCr*, *SIr*, *CIr and MIr* of region $R_i$. For every region, FINNs receives a total of 5-dimensional input data, and outputs the class label: *PR* and *NPR*. Then FINNs adjusts the center value and width of its fuzzy membership function automatically during the learning phase.

The bell-shaped membership function represents the if-part of fuzzy rules, which is placed between the input node *i* and the node *j* on the rule-layer. The membership function is expressed as

$$u_{ij} = \exp(-\frac{(F_i - w_{ij})^2}{\sigma_{ij}^2}), i = (1,2,...N_1); \quad j = (1,2,...,N_2) \qquad (7)$$

where $w_{ij}$ is the center value of the membership function, $\sigma_{ij}$ indicates the width of the membership function adjusted by the learning process in FINNs. In the rule-layer, degree of the *j-th* rule $\rho_j$ is computed as the following formula:

$$\rho_j = \min(\mu_{1j,}\mu_{2j,}...,\mu_{Nj}) \qquad (8)$$

And then, the estimated output node value $\hat{y}_k$ is calculated by the following equation:

$$\hat{y}_k = \frac{\sum_j^{N_2}(w_{jk}\rho_j)}{\sum_j^{N_2}\rho_j} \qquad (9)$$

where $w_{jk}$ is the weight between the *j-th* node in the rule-layer and the *k-th* output node. The logical form of the fuzzy inference if-then rules is given as

**If** $f_1$ is $\tilde{w}_{1j}$ and $f_2$ is $\tilde{w}_{2j}$, …, and $f_N$ is $\tilde{w}_{Nj}$ **then** $\hat{y}_k$ is $w_{jk}$, where $\tilde{w}_{ij}$ means the value near $w_{ij}$ depended on the value of $\sigma_{ij}$.

## 4.2 Learning process of FINNs

The learning process of the FINNs consists of the three phases. First, the center values of membership functions which correspond to the if-part and estimated values which correspond to the then-part are determined temporarily in the self-organizing phase. And then, the system merger the similar rules at the rule extraction phase. Finally, Least Mean Square (LMS) is executed to reduce the total mean-square error of the network to finely adjust the weights and the shapes of the membership functions.

### 4.2.1 Self-organizing Learning phase

In this phase, Kohonen's self-organizing algorithm is applied to the following two purposes. The first purpose is to estimate the center of membership functions of pre-condition part and the estimated value of *j-th* rule. The second purpose is to construct fuzzy if-then rules. In our implementation, the self-organizing learning phase and the LMS phase are almost the same as that of FINNs in [12].

### 4.2.2 Rule-Extracting Phase

In order to get better generalization ability, we employ a relative entropy-based approach to measure similarity between two rules described below.

For two probability densities functions $p_k$ and $q_k$, the Kullback-Leibler divergence is defined as

$$D_{p_k\|q_k} = \sum_k p_k \log\left(\frac{p_k}{q_k}\right) \qquad (10)$$

the Kullback-Leibler divergence indicates how distinguishable $p_k$ is from $q_k$ by maximum likelihood hypothesis testing when the actual data obeys $p_x$. It is well know that $D_{p_k\|q_k}$ is a nonnegative, additive but not symmetric. To obtain a symmetric measure, one can define similarity measure as:

$$SW(p_k,q_k) = \frac{D(p_k\|q_k)+D(q_k\|p_k)}{2} \qquad (11)$$

And then, for each weight vector $w_j$ ($j = 1,...,N_2$), we calculate a six-bin ($N_h = 6$) weight histogram $H_w(j,h)$ ($h = 1,...,N_h$), therefore, estimation of weight probability distribution function is calculated as

$$p_j(h) = \frac{H_w(j,h)}{\sum_{h=1}^{N_h} H_w(j,h)} \qquad (12)$$

Since small values indicate the densities are 'close', we merge two rules when $SW_{j,j+1}$ is smaller than the threshold $\delta_w$ which is selected by experiment described in the Figure 3. Therefore, the FINNs can combine rules to extract generalized rules, so as to improve generalization performance on pattern classification.

### 4.2.3 LMS Learning Phase

The goal of the LMS learning phase is to minimize the mean square error between outputs of the network and the desired signals, which can be achieved by adjust the parameters such as those to determine the shape and the center of the membership functions explained before. For the single-output FINN, the minimizing mean square error function is expressed as follows:

$$E = \sum_s E_s \qquad (13)$$

$$E_s = \frac{1}{2}(y-\hat{y})^2 \qquad (14)$$

where $y$ is the desired output and $\hat{y}$ is the output inferred by FINN. $s$ is learning pattern. According to the LMS learning principle, the estimation value of j-th node in the rule-layer is updated as

$$w_j^o(t+1) = w_j(t) + \varepsilon_{LMS}^w (y-\hat{y})\frac{\rho_j}{\sum_k^{Nr}\rho_k} \qquad (15)$$

The center and the width of membership functions are undated as

$$w_{ij}(t+1) = w_{ij}(t) + \varepsilon_{LMS}^{w}(y-\tilde{y}) \times \left(\frac{w_j^o \sum_k^{N_r} \rho_k - \sum_k^{Nr} w_k^o \rho_k}{\sum_k^{Nr} \rho_k}\right) \times q_{ij}\mu_{ij} \left. 2(f_i - w_{ij}) \middle/ \sigma_{ij}^2 \right. \quad (16)$$

and

$$\sigma_{ij}(t+1) = \sigma_{ij}(t) + \varepsilon_{LMS}^{\sigma}(y-\tilde{y}) \times \left(\frac{w_j^o \sum_k^{N_r} \rho_k - \sum_k^{Nr} w_k^o \rho_k}{\sum_k^{Nr} \rho_k}\right) \times q_{ij}\mu_{ij} \left. 2(f_i - w_{ij})^2 \middle/ \sigma_{ij}^3 \right. \quad (17)$$

respectively, where $q_{ij} = \begin{cases} 1 & \text{if } \rho_j = \mu_{ij}; \\ 0 & \text{elsewhere} \end{cases}$

## 5 Experimental Results

The proposed algorithm was integrated into a system that was tested using several video sequences. Table 1 summarizes the structural parameters of FINNs.

**Table.1** Structure of FINNs

| | |
|---|---|
| $N_1$ | 5 |
| $N_3$ | 2 |
| $\varepsilon_{self}(t=0)$ | 0.5 |
| $\varepsilon_{LMS}$ | 0.001 |
| $\sigma(t=0)$ | 4 |

In order to perform training, we randomly select three video sequences including 1152 frames. Rather than processing every frame of the video clip, our technique analyses only one out of every N (N=10) frames since there is typically little difference among consecutive frames. For each sampled frame, we label the ground truth manually; hence, the training data is composed of 117 sampled frames. As the result of the learning, the FINNs extracted 36 fuzzy inference rules. Since no perceptual theory exists regarding how to select the threshold $\delta_w$ in rules extraction phase, the parameter is determined by experimental tests illuminated as Fig.3. The results obtained from three digital video clips taken from a sports video*(Clip1)*, two movies: *"Season I of Friends"(Clip2)* and *"Great Forrest Gump"(Clip3)*. Due to the very strong subjectivity of human visual attention, the precision of our method is subjectively examined by ten testers and is averaged. Figure 3 shows the average precision curve with the use of different threshold $\delta_w$, as we can see more reasonable results of our method c ould be achieved by setting $\delta_w$ =0.15.

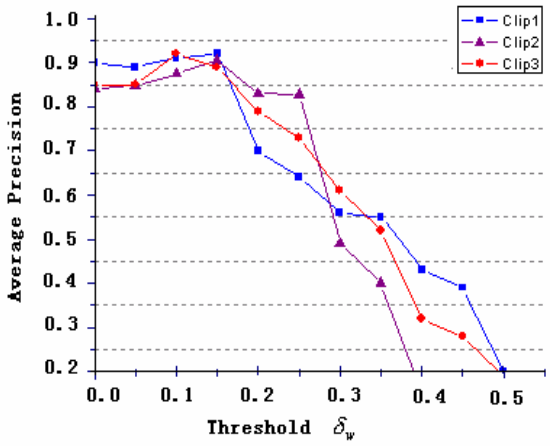



**Fig.3.** Average precision related to the estimated threshold $\delta_w$

The examples of rules obtained from the proposed system are shown in Table 2. These extracted rules are natural and are considered to be correct. The width of each membership function corresponds to the diversity of the input of the fuzzy rule. When the width of membership function is narrow, the input value is sensitive and has a large effect on the results. On the other hand, when the width is large, it means that the input is not very important. Therefore, we can estimate the importance of each input.

**Table.2.** Examples of extracted rules

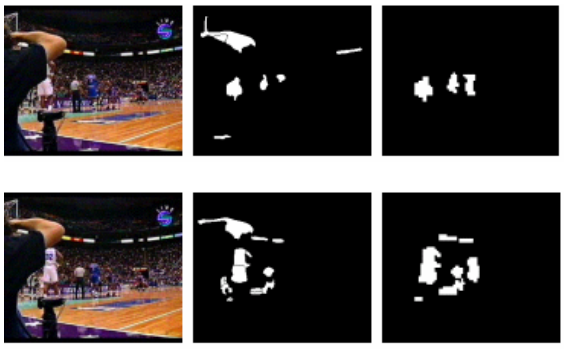
| No. Rules | $w_{ij}$ | $\sigma_{ij}$ |
|---|---|---|
| $R_1$ | $F_1$: 0.80 | 0.14 |
| | $F_2$: 0.67 | 0.06 |
| | $F_3$: 0.82 | 0.21 |
| | $F_4$: 0.33 | 0.29 |
| | $F_5$: 0.56 | 0.12 |
| $R_2$ | $F_1$: 0.66 | 0.12 |
| | $F_2$: 0.83 | 0.06 |
| | $F_3$: 0.67 | 0.20 |
| | $F_4$: 0.42 | 0.15 |
| | $F_5$: 0.67 | 0.09 |
| $R_3$ | $F_1$: 0.99 | 0.04 |
| | $F_2$: 0.74 | 0.08 |
| | $F_3$: 0.82 | 0.16 |
| | $F_4$: 0.52 | 0.12 |
| | $F_5$: 0.79 | 0.08 |

Experiment shows input features *CSr*, *OCr*, *Mir* have more important than other features, which are considered to be sound. Colour and motion have been found to be two of the strongest influences on visual attention [20], especially, a strong influence occurs when the colour of a region is distinct from the colour of its background. And our peripheral vision is highly tuned to detection changes in motion.

**(a) (b) (c)**

**Fig.4.** Prominent region detection in video sequence from *Season I of Friend*s

Fig.4 shows the results for two successive sampled frames taken from movie *Season I of Friend*s. Specifically, Fig.4 (a) shows the original frames, (b) gives corresponding results of homogenous region segmentation, and (c) shows prominent region detection results. Prominent regions are located and used to obtain a binary image that contains white and black pixels to represent prominent region and non-prominent regions, respectively. As shown in the fig.4, one of the background regions not only has high color contrast but also locates at near the center of image, so both of this region and the character Chandler draw more attention and have high perceptual prominence, which are correctly extracted.



(a) (b) (c)

**Fig.5.** Prominent region detection results for two frames taken from sports video clip: (a) Original image; (b)Detection results using low-level information; (c) Our detectioin results

Although a direct precision comparison with other system is not possible due to the lack of standard system setups and video sequences, we compared our results for a sports video clip with the results of the algorithm using low-level information (luminance, color, texture) described in [16]. Fig.5 shows the comparison results. As we can see, some noisy areas in Fig.5(b) are removed correctly. Our results also accord to the fact of human selective attention, namely when viewing the video clip, human will put little attention on these noisy areas, even though they have a high contrast. That is different from viewing single image.

In our experiments, several limitations are found. One major problem is caused by the noise areas, which have the same contrast and similar motion consistency as the prominent regions. As demonstrated in Fig.6, one of background regions is assigned mistakenly a high perceptual prominence. However, we expect this drawback can be improved by using spatial relations analysis, which is one of our future works. The shadow region of object is the other limitations of our system, which is difficult to handle because, in many cases, shadow region may locate at the center of an image, mean that it has higher value of compositional balance and color contrast. In Fig.6, the shadow region between two characters is regarded as a prominent region mistakenly.

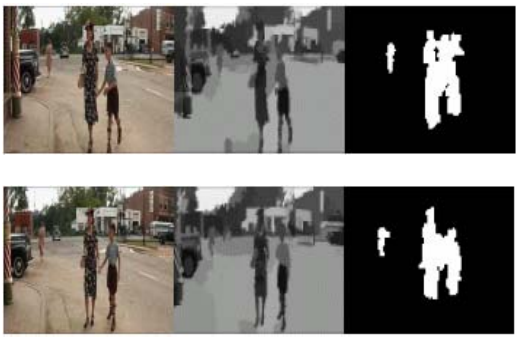


**Fig.6.** Prominent region detection in video sequence from "*Great Forrest Gump*"

# 6 Conclusions

A perception-oriented approach for identifying prominent region in video sequences is presented in this paper. We extract a number of different perceptual features by the taking advantage of the *mise-en-scene* principles, which is different from many previous researchers who have used only low-level features. Furthermore, considering the subjectivity and imprecise of human perception, a modified fuzzy inference neural networks is ideal for classifying prominent regions due to the combination of learning ability of neural networks and rule processing ability of fuzzy logic. We provide a mechanism for prominent region detection through soft decisions. This framework can adequately capture subjectivity involved in the perceptual promience of region. And then, the technique has been designed to easily accommodate application specific requirements. Although the experimental results show the encouraging performance, the conducted research also shows that there is plenty of room for improvement.

Future work will be focused on how to handle the limitations in the approach and improve the results. Additional research work includes a suitable pattern description for the video sequences with the use of the extracted prominent regions.

## References

[1] J.Fan, W.G.Aref, A.K.Elmagamid, M.S.Hacid, M.S.Marzouk, and X.Zhu: Multi View: Multi-level Video Content Representation and Retrieval. J.Electron.Imaging ,special issue on multimedia database 10(4) (2001) 895-908

[2] Y.Deng and B.S.Majunath: NeTra-V:Toward an Object-based Video Representation. IEEE Trans. Circuits Syst. Video Technol., 8 (1998) 616-627

[3] S.F.Chang, W.Chen, H.J.Meng, H.Sundaram and D.Zhong: A Fully Automatic Content-based Video Search Engine Supporting Spatiotemporal Queries. IEEE Trans. Circuits Syst. Video Technol., 8 (1998) 602-615

[4] J. Senders: Distribution of Attention in Static and Dynamic Scenes. In: proceedings SPIE 3026 (1997) 186-194

[5] A. Yarbus: Eye Movements and Vision. Plenum Press, NewYork NY, (1967)

[6] L.Itti, C.Koch: Feature Combination Strategies for Saliency-based Visual Attention Systems. Journal of Electronic Imaging, 10(1) (2001) 161-169

[7] D.Parkhurst, K.Law, and E.Niebur: Modeling the Role of Salience in the Allocation of Overt Visual Attention. In: proceedings ICIP, (2003)

[8] David Bordwell, Kristin Thompson: Film Art: An Introduction. McGraw-Hill Higher Education, (2001)

[9] D.Comaniciu, P.Meer: Mean Shift: A Robust Approach toward Feature Space Analysis. In: IEEE Trans. Pattern Analysis Machine Intelligence, 24 (2002) 603-619

[10] S.Marcelja: Mathematical Description of the Responses of Simple Cortical Cells. Journal of Optical Society of America, 70 (1980) 1169-1179

[11] Y.F.Ma, H.J.Zhang: A Model of Motion Attention for Video Skimming. In: proceedings. ICIP (2002) 22-25

[12] T.Nishina, M.Hagiwara: Fuzzy Inference Neural Network. Neurocomputing, 14(1997) 223-239

[13] H.lyatomi, M.Hagiwara: Scenery Image Recognition and Interpretation Using Fuzzy Inference Neural Networks. Pattern Recognition 35(8) (2002) 1793-1806

[14] Jia Li, James ze wang and G.Wiederhold: Simplicity: Semantics-Sensitive Integrated Matching for Picture Libraries. In: IEEE Trans. Pattern Analysis Machine Intelligence, 23(9) (2001)

[15] Ying Li, Y.F. Ma and H.J.Zhang: Salient Region Detection and Tracking in Video. In: Proceedings of ICME (2003) 269-272

[16] Alexander Dimai: Unsupervised Extraction of Salient Region-Descriptors for Content Based Image Retrieval. In: Proceedings ICIAP (1999) 686-672

[17] Lotfi A.Zadeh: A Note on Web intelligence, World Knowledge and Fuzzy Logic. Data&Knowledge Engineering, 50 (2004) 291-304

[18] E.J.Pauwels, G.Frederix: Finding Salient Regions in Images Nonparametric Clustering for Image Segmentation and Grouping. Computer Vision and Image Understanding 75 (1999)

[19] Hang Fai Lau, Martin D.Levine: Finding a small number of regions in an image using low-level features. Pattern Recognition 35 (2002)

[20] E.Niebur and C.Koch. Computational architectures for Attention. In R.Parasuraman, The Attentive Brain. MIT Press (1997)