www.s2is.org

International Journal on Smart Sensing and Intelligent Systems

ISSN 1178-5608

APTIVE IMAGE SEGMENTATION BASED ON SALIENCY DETECTION

Shui Linlin

School of Animation and Digital Media Art， Communication University of China

Beijing 100024, China

Submitted: Oct. 3, 2014 ***Accepted: Jan. 20, 2015*** ***Published: Mar. 1, 2015***

Abstract- in this article, we propose an adaptive image segmentation method based on saliency. First of all, we obtain the saliency map of an image via four bottom-layer feature tunnels, i.e. color, intensity, direction and energy. The energy tunnel helps to describe the outline of objects better in the saliency map. Then, we construct the target detection masks according to the greyness of pixels in the saliency map. Each mask is applied to the original image as the result of pre-segmentation, then corresponding image entropy is calculated. Predict the expected entropy according to maximum entropy criteria and select the optimal segmentation according to the entropies of pre-segmented images and the expected entropy. A large number of experiments have proved the effectiveness and advantages of this algorithm..

Index terms***:*** **Visual feature, image segment, maximum energy criteria, saliency detection**

I. INTRODUCTION

Driverless technique is developing ceaselessly. In 2010, a group of 7 Google driverless cars were put into trial in California. Google driverless cars are issued with the first driving permit by DMV in Nevada, USA, which means that they can operate in this state. In 2011, the driverless technique developed by National University of Defense Technology is successfully to domestic vehicles and a driverless car traveled in structuring road for 177 miles, from Changsha to Wuhan. Alan Taub，the vise director in global research apartment of Motor General, predicted that in 2020 the driverless system would be applied to every vehicle. Driverless technique has gradually permeated into our life and is the trade of future development. Meanwhile, although related cars are allowed on road, they are not popularized, which shows that such technique is not applicable on all conditions. Google driverless cars can drive in Nevada, but it is not sure whether they can travel in other states or the more crowded roads in Beijing. Anyway, driverless cars cannot avoid to 'see' with their 'eyes' and perceive the various objects via 'vision'.

In driverless technique, target detection is like the visual system of intelligent driving controller. It performs target detection, classification, tracking, location and detailed segmentation via techniques such as signal processing, computer vision, image processing, machine learning, patter recognition, artificial intelligence, data mining, and multimedia retrieval. Target detection uses multi-information fusion technology，in which cameras are indispensable. This shows the importance and necessity of real-time video processing. Compared to radar, the advantages of real-time video processing are irreplaceable. Typically radar can only locate the target, but it cannot detect the content, so some meaningful targets, such as traffic marks, cannot be detected by radar. Image processing can analysis the content as well as location of targets, however the present image information is too complicate for fast processing, so the utilization of information fusion can improve the efficiency. In the future, with the development of IT, update of intelligence algorithms and improvement of computers, it is believed that target detection by visual systems can replace radars. The salient features of vision are originally obtained from human visual cognitive system. With the development of psychology and physiology, researchers gradually get to know the visual systems of Primates and convey the mechanisms of psychology and physiology using information sciences. Thus, we have abandoned target analysis from the angle of image. Rather, we start from human visual system and applied visual features to target detection and analysis, which is more reasonable. Therefore, target detection based on visual salient features is critical and has drawn the attention of many researchers.

For the time being, target detection is widely applicable through various application techniques. It can be realized via infrared radar, the features in image as well as multi-information fusion. In this article we focus on detect targets in images. The typical method of target detection in images is target analysis via features in the images, thus image features are important data. The classification of target detection methods is complicated. At the angle of targets, it can be sorted into dynamic target detection and static target detection. From the angle of image features, it can be classified into target detection based on local features and target detection based on global features. Analyzing images from the view of human visual cognition, there are some regions with relative strong stimulus to human vision. Target detection can also be realized through visual saliency.

Target detection based on local features usually describes the target via the key points of the target and the image as well as their neighboring regions, or through the features in some regions. The research on local features started from late 1970s.

Gao[1] proposed two multi-scales corner detection methods based on log-Gabor WT, which were LGWTOI method and the LGWTSMM method. The LGWTOI method was simple and obtains some extra information which can be used in the following processing. To achieve isotropic response, the LGWTSMM method exploied the multi-orientation decomposition of log-Gabor WT and constructs the second moment matrix. The proposed methods provided a unique response for the corners of higher order structures. The authors[2]proposed an improved multi-scale corner detector based on Curvature Scale Space (CSS) technique. The proposed method offered a robust and effective solution to images containing widely different size features. In the paper [3], Harris used moving bright block judgment instead of differential operators, which was more accurate for corner detection and robust to rotational and grayness variance. Harris Corner Detection is still used today. Combining Gauss Scale Space with Harris Corner detection algorithm and using Iterative Estimation of Lindeberg [4-6] to realize affine invariance within the neighboring region, Harris-Laplacian Detector makes sure that Harris-Laplacian Corner is scale invariant. Harris-Affine Detector can automatically detect image features under affine transformation to adapt to such changes. In 2004 Lowe proposed the famous SIFT local feature [7, 8], whose full name is Scale Invariant Feature Transform. This feature is adaptive to changes of scales, orientation, view angle and affine changes. The algorithm detects extreme points in Gauss-Laplace Space as key points via the difference between Image Pyramid and Gauss Kernel Filter and used a 128-dimensional vector for feature description, which makes this method more adaptive to applications. In 2006 Bay followed the idea of SIFT and proposed Speeded Up Robust Feature, or SURF for short. This method speeds up key point detection via combination of Harr Wavelet and Integral Image. SIFT algorithm is frequently improved afterwards. People have added other features besides greyness to make the matching of local features more accurate. In case of excessive key points, some researchers put forward global DAISY as the local feature descriptor which is fast to calculate.

Detection can be based on target structures. The structure of a target can reflect its properties, so detecting it as a feature of the target can realize accurate detection. Typically objects consist of structured features, for example a human consists of the head, the body and the four limbs, a vehicle of the body and the tires and a face of five sense organs. Such structured information can help to accurately detect targets from over complicated background. Dalal, Triggs et al detected pedestrians using HOG and achieved outstanding results, and HOG is widely applied in feature expression. But single global template is limited that it cannot adapt to various gestures of pedestrians. Felzenszwalb et al used part-based model to present target during detection and meanwhile trained with Latent SVM classifier, which was proved more efficient than single template. Detection methods except for the one mentioned above is all based on local image features or their combinations. Target detection can also be conducted through the differences in global features. Detection methods based on global features include Background Difference and Successive Frame Difference. Background Difference is relatively easy to understand. It is usually used in fixed supervisory and control devices. Set the background unchanged and as a prior knowledge, if an object other than the background appears, generate the target via the difference between the target image and the background image. This method is easy to realize but requires fixed background. Researchers proposed multiple background parameter models, trying to reduce or eliminate the influence of changing background on moving target detection, or to reduce the complexity of calculation to meet the real-time need of the system. Successive Frame Difference is also a target detection method based on global feature difference.

Visual Saliency presents the intensity of visual stimulus sensed by the visual system and the saliency map is the direct presentation of Visual Saliency. Saliency map can be obtained from the combination of visual features through visual cognitive mechanisms. Saliency map constructed based on human visual system can provide guidance for fast location of salient visual region and highlight useful information during target detection. The calculation model of visual saliency map can distribute limited processing resources to a few salient regions. Visual saliency research starts from human visual cognition to analyze how humans use our visual systems to observe the word. Though physiology, psychology and eye trackers, we probe into how eyes perceive information, which visual information causes strong stimulus to the visual system, how higher-level information is obtained via processing the information by the brain, through what mechanisms higher-level information is obtained, what general laws are there in visual cognition and how to use these mechanisms as well as laws to construct more useful information from bottom-level features. In static target detection, we propose adaptive salient target segmentation method to fast separate targets and the complicated background and effectively detect salient targets from images. This method is developed from Visual Attention Model and Maximum Entropy Segmentation algorithm. First of all, we obtain the saliency map of an image via four bottom-layer feature tunnels, i.e. color, intensity, direction and energy. The energy tunnel helps to describe the outline of objects better in the saliency map. Then, we construct the target detection masks according to the grayness of pixels in the saliency map. Each mask is applied to the original image as the result of pre-segmentation, then corresponding image entropy is calculated. Predict the expected entropy according to maximum entropy criteria and select the optimal segmentation according to the entropies of pre-segmented images and the expected entropy. Experiments have proved the effectiveness of this algorithm.

II. RELATED KNOWLEDGE

A. visual features

Generally, abundant information is contained in a visual scene, including vision, hearing, touch and taste. Humans usually perform synthetic analysis to all kinds of information and obtained the needed part. But the visual system carries more tasks of information processing during cognition, so researches on vision are critical in scene comprehension. Visual Salient Features are image features considered from the visual cognition of humans or higher animals. Images are regarded as stimulus to vision and converted into visual salient features according to visual cognitive laws based on physiological and psychological experiments. Primary visual features were firstly mentioned in visual attention mechanisms, which is a bottom-up mechanism.

Under this mechanism, creatures with visual ability will quickly select and concentrate their vision on a region of an image. This quick selection and concentration is based on certain rules. This extraneous stimulus that influences on visual attention is the primary visual features in visual information. Primary visual features are various and complex and those that have been proved by experiments to have influences on pre-attention are direction, color, scale, curvature, depth, movement, measurement shape and so on. Visual saliency is the contrast formed by some visual stimulus brought by visual objects. Such visual stimulus originates from feature description, and the extraction of visual salient features is to process the primary visual features which can stimulate vision. Thus, we cannot perceive the world only by primary visual features, rather we need complex mechanisms to deeply process them. Common visual features are color, brightness, and direction, which are critical for the bottom-up attention mechanism and can effectively describe an image.

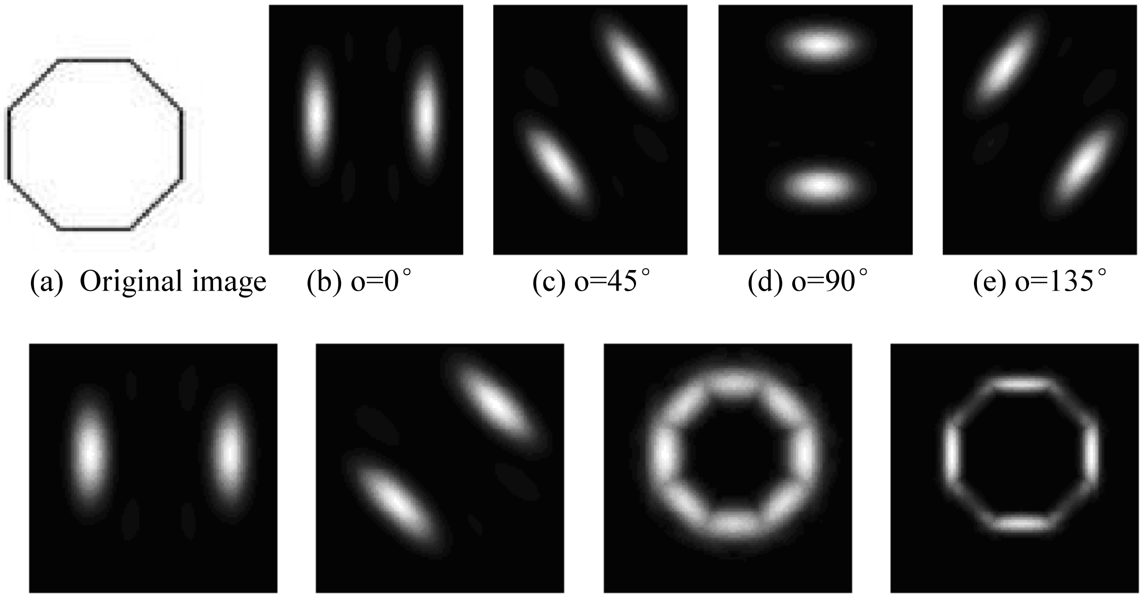
1) Direction

Direction is independent of color and brightness and reflects the information of an image alike. Gabor function is regarded as a mathematic description of a kind of simple cells of the visual cortex. In 1980, Marcelja discovered that 3-dimensional Gabor function is the optimal descriptive function for the sensory fields of simple cells. Later in the area of neuropsychology, Pollen and Ronner found via experiments on the cells of the visual cortex of cats that the two-dimensional image of adjacent cells was similar to Gabor function, and this kind of cells were even symmetric or odd symmetric which is also shown in Gabor function. Finally in electrophysiological experiments, Jones and Palmer discovered through testing the spatial spectrum of simple cells on visual cortex that the response properties of simple cells were similar to Gabor function.

Two-dimensional Gabor function is:

$$g_{\sigma,w_0}\left(x,z\right)=\frac{1}{2\pi\sigma_x\sigma_y}e^{-\frac{1}{2}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2})}e^{j\omega_o(x+y)} \qquad (1)$$

The parameters σ_x and σ_y is the standard deviation of Gauss function in the direction of x and y. ω_0 is the spatial frequency. Figure (1) is the original image in the experiment and λ, the wavelength of two-dimensional Gabor function, is set to 8. Subfigure (b), (c), (d), (e) are Gabor function with θ, the direction parameter, set to 0^o, 45^o, 90^o and 135^o. Subfigure (f) and (g) is the image of $\theta= 180^o$ and $\theta=225^o$ after filtration. It can be concluded that on the image spatial field, the filtered fusion images are the same on the direction of 0^o and 180^o, 45^o and 225^o. Thus, the range of direction parameter θ is [0^o 180^o]. Subfigure (h) is the fusion image of filtration with directions of $\theta=0^o$, 5^o, 90^o and 135^o. Subfigure (i) is the fusion image of filtration with directions of $\theta=0^o$, 30^o, 60^o, 120^o and 150^o. It is shown that the more directions there are, the more accurate the texture of an image is depicted and similarly the more detailed the directions are reflected.



(a) Original image (b) o=0° (c) o=45° (d) o=90° (e) o=135°

(f) o=180° (g) o=225° (h) feature fusion map with four directions (i) feature fusion map with six directions

Figure 1 Direction function performance verification

2) Color

Color is the subjective sensation of human brains. There is no reliable result on color analysis via neurological mechanisms. If the stimulus is only caused by red, green and blue, we cannot perceive the color, but only the brightness. When entering a dark environment from a bright one, the sensations of vision under bright and dark conditions are carried out by different cells. In bright environment, the sensory cells are cone cells, otherwise they are rod cells. Visual sensory fields response according to the wavelength of different lights. In bright conditions, human eyes can clearly distinguish more colors on the spectrum, while when the brightness decreases to a certain level, we can only tell the change of darkness and brightness and lose the discriminative ability to different colors. RGB (red, green, blue) color space is also called computer graph color space, which converted the colors of images into information expressible by computers. It is a widely used and the most fundamental color space, and all present color spaces can be converted to RGB color space. Although RGB is easy to present by computers, it is not consistent with the human cognition. HSI (hue, saturation, intensity) color space is constructed based on human visual system, and it describes the colors of images via hue, saturation (Chroma) and intensity (brightness). HSI color space is presented by a cone model. Although the cone model is complicated in description, it can clearly present the changes of hue, saturation and intensity. Hue and saturation are called by a joint name, tone. Tone usually presents the type and shade of colors. Humans are more sensitive to brightness than shade, so according to the requirements of color processing and cognition, HSI color space, which is consistent with human visual system, is more coincident with the properties of human visual cognition than RGB color space. At present, many algorithms in computer vision can be used in HSI color space, which can process hue, saturation and intensity separately and keep their independence from each other. Therefore, HSI color space is highly practical.

3) Opponent Color Theory

In the mid-19th century, Heirng proposed Opponent Color Theory, which states that there are opposing visual sensations between red and green, yellow and blue, as well as black and white. Researchers commonly believe that in human light receptor, color information is received by red, green and blue cones and transmitted along nerves in the form of color contrast. Opponent color is expresses in detail as:

$$RG = R - B \tag{2}$$

$$BY = B - (R + G) / 2 \tag{3}$$

Hence RG is the contrast component of red and green and BY is that of blue and yellow.

4) Brightness

From the analysis of the color we can see the existence of certain relations between brightness and color. In fact, brightness describes an image by means of grayness. It is a merit in optic and also reflects human psychological physical merit. Suppose we define some merits to present the amount of radiation in standard conditions, the relations between radiation and light sensation can be constructed. The range of wavelength for visible light of 380~780mm, the physical merits given by radiation within is range is visible to humans and converted to psychological physical merits. From the angle of cognition, brightness is the measurement of the intensity and energy of the light that a visual receptor receives, so it is a subjective measurement of the amount of radiation. If the image is chromatic, we can convert it to a grayness one via an equation. However, this process is not reversible, i.e. we cannot convert a grayness image to a chromatic one. The equation changing an image in RGB color space to a grayness one is:

$$I = 0.299 \times R + 0.587 \times G + 0.114 \times B \tag{4}$$

Therein I is the brightness and R, G and B respectively present red, green and blue components in the color space.

B. Visual cognition model

This article mainly analyzes global features, constructs visual cognition model via researches on visual salient features and visual cognition mechanism, and realizes target detection. The idea is to search for the significant difference via multi-feature expression of images. Using the differences of multiple features or time requires an effective combination of features to obtain the salient feature image. For the time being, Itti model is a classic model among visual salient feature models. It combines multiple primary visual features via visual cognition mechanisms and is applied to visual attention, simulating the behavior of visual attention via saliency map. With the successful example, researchers have continued to enrich and probe this research based on Itti Saliency Map model. They have added to the construction methods of saliency map via different theories such as Graph Theory and statistics, and employed them in various areas, including scaling based on image saliency, target detection and location, image retrieval and computer vision. This article is based on feature expression method in living visual cognition as well as representative algorithms such as Itti model and HMAX model. We started from bottom visual features and effectively apply visual salient features into Selective Attention Algorithm (Itti Algorithm) via simulation of the visual attention mechanisms in animals. Itti model and HMAX model are classic visual models, both of which are based on visual physiological mechanisms. HMAX model was first put forward by Riesenbuber and Poggio, and Mike Tarr simplified its full name Hierarchical Model and X to HMAX.

There are two secondary vision tunnels, one is Ventral Pathway, also named Ventral Stream， and the other is Dorsal Pathway, also called Dorsal Stream. Ventral Stream is charge of the detection of objects and Dorsal Stream is in charge of analyzing their spatial and topological information. HMAX model is based on Ventral Stream and combined with the physiological properties of Dorsal Stream. It constructs visual cognition model using information science manner according to the physiological sequences that different cortices function when the brain is detecting objects. In this model, data from Inferior Temporal Cortex of macaques, i.e. the information of neurons of different shapes and properties in the secondary visual regions of Ventral Stream, are recorded using Logothetis. In brain cortices, target detection is conducted by Ventral Stream. It starts from primary visual cortex V1, goes through texture cortex V2 and V4, and finally ends in Inferior Temporal Cortex. Researchers discovered through experiments on macaques that Inferior Temporal Cortex (IT) is the primary cortex for target detection in higher animals. In addition, this cortex provides information for Recognition and Control Center （PCF） and participates in the process of target cognition and recollection.

For years researchers have performed physiological experiments on Primates and achieved a number of cognitive structure theories. They have constructed and realized related cognition mechanisms according to these theories and proved or extended them via physiological experiments, which makes this discipline popular. The series of researches started from physiological experiments on macaques and cats by Hubel and Wiesel, went deep to the exploration on V1 and simple cells, and reached to the fact that the secondary region of Ventral Pathway only responses to the complex target stimulus. The structural properties of Ventral Pathway are summarized as the following: 1. It is invariant to structure levels and robust to the complex disturbance of spatial locations, scales and changes in view angle; 2. The sensory fields for targets increase gradually; 3. It performs basic feed forward information processing for target detection; 4. It has certain active learning capacity in IT cortex; 5. It can maintain the scale and position during single target detection; 6. With the deep-going of neuron learning, the information will gradually become complicated to reach the optimal neuron stimulus.

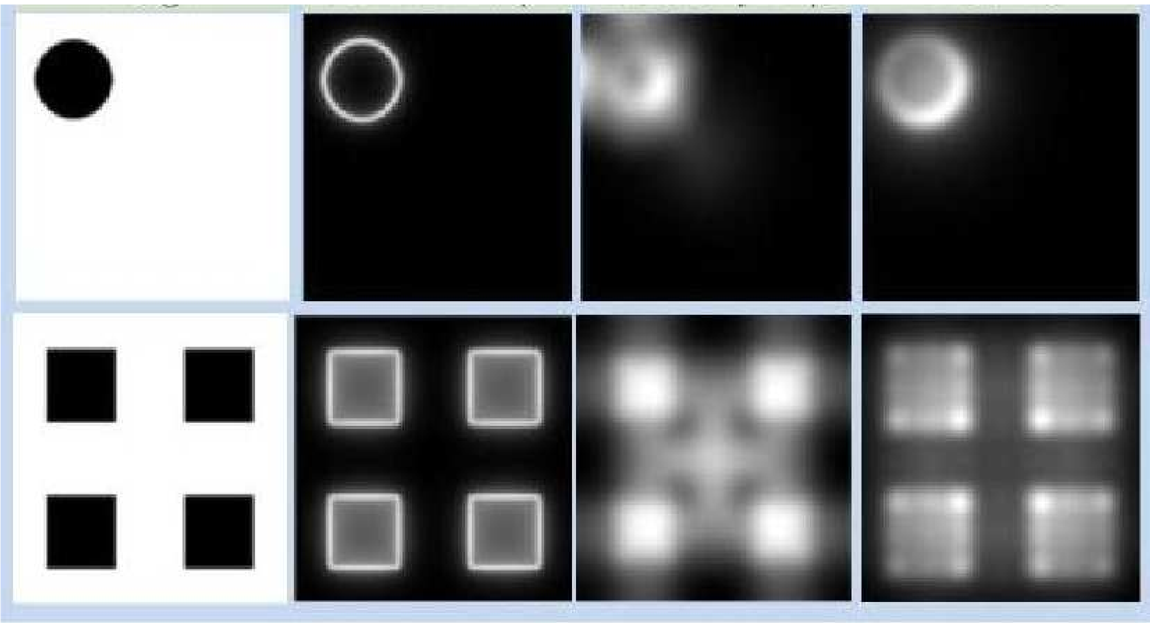
III. ADAPTIVE TARGET DETECTION AND SEGMENTATION BASED ON SALIENCY

Based on the former analysis on features and their combinations as well as the requirements for target detection, we select proper target features and realized automatic image segmentation via maximum entropy method. Salient target detection methods, whether based on physiology or mathematics, place high importance on the consideration of saliency from local differences in images and the independent completeness of targets. This article starts from the angle of physiology and added the local energy tunnel to the traditional three-tunnel model. We obtain the local feature difference via color, intensity and direction tunnels, and get the outline information from the energy tunnel to enhance the global independence of salient objects. The saliency of objects is determined automatically via maximum entropy criteria. Thus, we have proposed a target detection method which is consistent with visual physiological properties and able to adaptively separate the background.

A. Saliency map of fusion local energy

First of all we construct the saliency map. The uniform features of Scene Image $f(x)$ in different scales are extracted via the parallel tunnel of color, direction, intensity and local energy pyramids. Then they are normalized and fused into the feature graph under different scales to get the saliency map of the scene image.

Classical Itti Model employs Gabor function as the direction tunnel. This function is consistent with the structure of sensory field cells and responses effectively to the direction information in images. Gabor function has a strong response to the central region of an image and a weak one to the non-central regions. This property makes it prone to lose the feature information of targets away from the center under complicated backgrounds, while the inclusion of local energy can payoff this shortcoming.



(a) (b) (c) (d)

Figure 2 Experimental results of different methods
((a) is the original image; (b) is the local energy figure; (c) is the three-tunnel salient map; (d) is the salient map of our method)

Local energy is positive correlated to the consistence of phase position, so the introduction of local energy as a saliency tunnel makes up for the shortcoming of sparse and discrete consistency feature as well as contains the effective response of consistency to the whole map. In Figure 2, the first column is the original image, the second is the local energy figure, the third is the three-tunnel (color, direction and intensity) saliency map and the fourth is the four-tunnel saliency where the energy tunnel is added. It is shown that local energy can response effective to both the central and peripheral regions, and the fusion with local energy tunnel weakens the central intensification proper of Gabor function, which reinforces targets in the peripheral region and makes sure the completeness of target detection. Neurophysiological experiments showed that the early stage of visual processing is dominant by the processing of edge features, or line feature. Andrea Perna et al [9] used un-processed edges, phased changed edges and outlines with noises to stimulate Primary Visual Cortex and found out that phased changed edges put the strongest stimulus to the cortex, which implied that phase consistency is coded by Primary Visual Cortex. Although phase consistency can effectively describe edges, during the experiments the image features were hard to describe after multiple tunnel fusion because of the sparseness of edge information. The authors [10] discovered that local energy contained in points with consistent phase was concentrated, and local energy is an important property of visual cortex cognition. Similar to phase consistency, local energy also has effective response to features of the whole image, and it also makes up to makes up for the shortcoming of sparseness of phase consistency. Local energy model is constructed from the odd items and even items of Gabor function. Venkatesh and Owens gave the complex expression of local energy:

$$E(x) = \sqrt{I^2(x) + H^2(x)} \tag{5}$$

In this function, $I(x)$ is the real component and $H(x)$ is the imaginary component. In addition, the relation between phase consistency and energy is:

$$E(x) = PC(x)\sum A_n \tag{6}$$

Here in $E(x)$ is the local energy, $PC(x)$ is phase consistency, and An is local amplitude.

To address the low speed of Hilbert Transformation, in 2001 Micheal Felsberg and Gerald Sommer from Germany used Riesz Transformation to construct the local energy function [12] and solved the problem caused by Hilber Transformation. The Kernel function of Risez Transformation is:

$$H(u) = -i\frac{u}{|u|} \tag{7}$$

In this function, u is the variable within the frequency band of signal analysis and this kernel function functions within the frequency band. This signal analysis method is not feasible to apply directly to image, because it requires generation of kernel template and the above function above:

$$h(x) = \frac{x}{2\pi |x|^3} \tag{8}$$

In this function, x is the range of the image and $f(x)$ is the original signal. Signal after Riesz Transformation is called Monogenic Signal, which is represented by $F_M(x)$:

$$f_M(x) = f(x) - h(x) \times f(x) \tag{9}$$

To find the phase information of different frequency band, Log-Gabor Transformation is used as scalar variable, which equips the scalar variable in Riesz Transformation Space with wider bandwidth. Riesz Kernel after improvement is [13]:

$$H(u) = -i\frac{u}{|u|}G(|u|) \tag{10}$$

Herein $G(|u|)$ is a one-dimensional Log-Gabor function. Coordinates in two-dimensional space is shown as:

$$\left\langle \begin{aligned} H(u_1) &= -i\frac{u_1}{\sqrt{u_1^2+u_2^2}}G(\sqrt{u_1^2+u_2^2}) \\ &= i\cos\theta G(\omega) = H_1(\theta,\omega) \\ H(u_2) &= -i\frac{u_2}{\sqrt{u_1^2+u_2^2}}G(\sqrt{u_1^2+u_2^2}) \\ &= i\sin\theta G(\omega) = H_2(\theta,\omega) \end{aligned} \right. \tag{11}$$

In the range of image, we use triple to construct Riesz Transformation Space:

$$\left\langle \begin{aligned} p(x) &= (f * G_x)(x) \\ q_1(x) &= (f * h_1)(x) \\ q_2(x) &= (f * h_2)(x) \end{aligned} \right. \tag{12}$$

$G(x)$ is the convolution of the image and one-dimensional Log-Gabor Transformation. h_1 and h_2 are the product of $H(u)$ and the convolution template h(x). Thus, the local amplitude of Monogenic Signal A(x) is shown as:

$$A_n(x) = \sqrt[2]{p^2 + q_1^2 + q_2^2} \tag{13}$$

The local energy is:

$$E(x) = \sqrt[2]{\sum A_n + \sum q_1^2 + \sum q_2^2} \tag{14}$$

B. Salient target segmentation based on Maximum Entropy Estimation

The idea of Maximum Entropy Estimation is inspired by maximum entropy segmentation. In Information Theory, entropy is used to measure the uncertainty of information and presents the average amount of information. Shannon put forward the definition of entropy:

$$H(A) = -\sum_{i=1}^{n} p(i)\log p(i) \tag{15}$$

Entropy Theory is also applicable on images. In the above formula, *p*(i) is the probability of corresponding degree of grayness and *H*(*A*) is the entropy of the image. Obtain the target entropy through Maximum Entropy Criteria [14, 17-18]. Theory of Maximum Entropy is usually used in image segmentation. It assumes that the changes in grayness are similar in target region and background region, and the fluctuation of grayness is smooth. When the amount of information is large both in target region and in background, the entropy of image reaches its maximum, and the corresponding grayness is the threshold of image segmentation. At this time, if the threshold value is increased or decreased, target region and background region will be confused during segmentation. When the entropy reaches the maximum, segmentation between background and targets reaches its optimal effect. At this time, the entropy in the target region is the amount of information. If the size of Image $f(x)$ is $M*N$ and the set of grayness is $G=\{0,1,\ldots,L-1\}$. Set n_i as the frequency of pixels whose grayness value is *I* and p_i is the corresponding probability:

$$P_O(t)=\sum_{i=0}^{t}p_i \tag{16}$$

$$P_B(t)=\sum_{i=0}^{t}p_i \tag{17}$$

And it is known that:

$$P_O(t)+P_B(t)=1 \tag{18}$$

At this time, then entropies of target region O and background region B are:

$$H_O(t)=In\sum_{i=0}^{t}\frac{p_i}{P_o(t)} \tag{19}$$

$$H_B(t)=In\sum_{i=t+1}^{L-1}\frac{p_i}{P_B(t)}=In\sum_{i=t+1}^{L-1}\frac{p_i}{1-P_o(t)} \tag{20}$$

The total energy is:

$$H(t)=H_O(t)+H_B(t) \tag{21}$$

When $H(t)$ reaches the maximum value, the value of $H_\mathrm{o}(t)$ is the entropy of target region.

In the last section we have obtained four-tunnel saliency map $g(x)$. Regions with intense grayness in $g(x)$ represents salient $f(x)$ feature in the original image. Sort $g(x)$ images into 99 orders according to the grayness. x_t is a point in x_i, and $g(x_\mathrm{t})\in g(x_\mathrm{i})$, and divide the saliency map into two parts $g(M)$ and $g(m)$, in which $g(M)>g(x_\mathrm{t})$ is the set of salient regions and $g(m)<g(x_\mathrm{t})$ is the set of non-salient regions. Let $g(M)=1$ and $g(m)=0$ so we can obtain the binary mask $g(t)$. The image of salient region is presented as:

$$f(M)=(f*g)(x) \tag{22}$$

Calculate the entropy of $f(M)$, $H(M)$. When $H(M)\approx H_0(t)$, $f(M)$ is the final target detection image. The first step is to construct image pyramid and calculate the target entropy according to Maximum Entropy Criteria. Then the saliency maps are obtained via color, intensity, direction and local energy tunnels, sort them according to saliency. Obtain the binary masks, apply them to the original image and calculate the corresponding entropy. The image whose entropy is the most approximate to the target entropy is the result of target detection.

IV. EXPERIEMNTS

A Experimental comparison to Itti visual model

. In this article, MSRA dataset [15] is selected as experimental image set and its remarks are used as the criteria to test the experimental results. In the first experiment, the parameters in saliency model are adjusted according to ROC curve. In the second experiment, the algorithm is subjectively evaluated according to the result of target detection, and the effectiveness of target detection is objectively evaluated according to the discovery rate, accuracy rate and F-measure [16].

Experiment 1: we selected five image samples and set the tunnel weights via ROC Curve, which is shown in Figure 2. Mainly the weights of direction and local energy tunnels need to be determined, and those of other tunnels are set to 1. In the local energy tunnel, we set the dimension to three, the minimum wave length to 3 and frequency amplitude to 2 so as to better describe the outlines of salient targets. It is shown in Figure 2 that when L, the weight of local energy tunnel, is set to 1.3, and G, the weight of direction tunnel, is set to 0.7, we can achieve the optimal visual saliency model. Figure 2 also shows that the area of ROC curve when L=1.3 and G=0.7 is larger than that when L=1.4, R=1.6 and when L=1.1, R=0.9, and the modified model with local energy tunnel is better than the tradition Itti model.

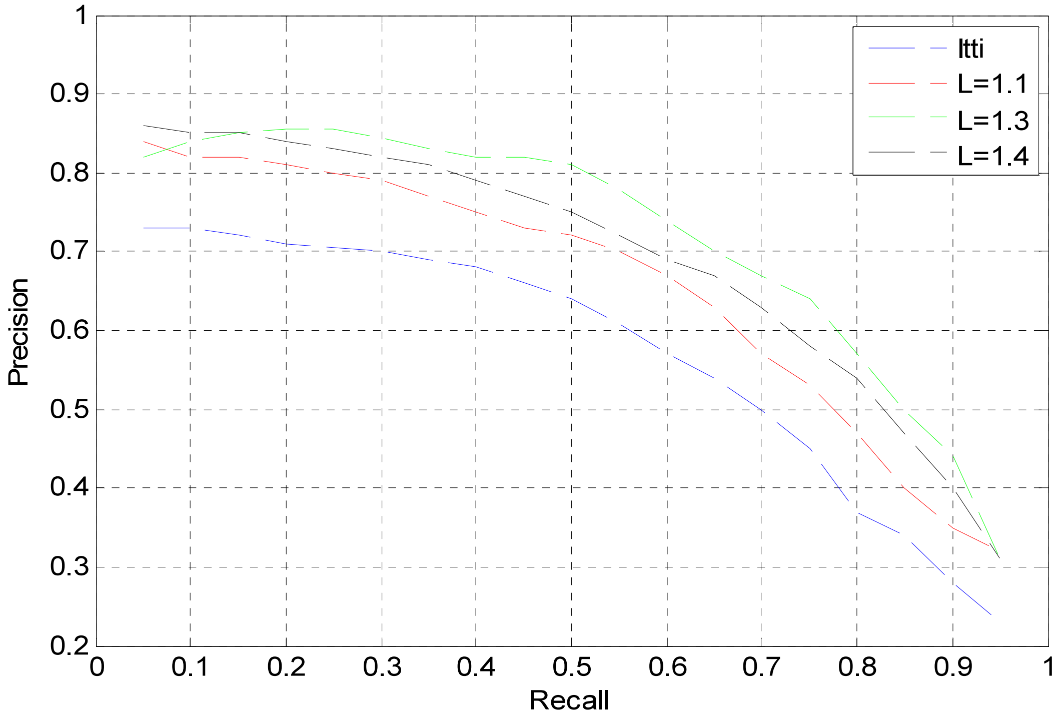


Figure 3 ROC Curves for selecting the weight of channel

Experiment 2: the comparison between out method and Itti. In Figure 3, the first column is the original image, the second is Itti saliency map, the third is salient target detection algorithm based on Itti Model, the fourth is the saliency map constructed with our methods and the last is the automatic target detection algorithm in this article. It is shown that our algorithm effective preserves the completeness of the target. In the image of the first and second row, the targets are located at the center of scene. From the target detection image we can see that in Itti model, part of the saliency in the duckling and mark is missing due to the lack of outline information. On the contrary, our algorithm preserves the completeness in these two images. Images in the third row are testing images for multi-target detection, and the targets are four squares located at four corner. Itti model is not able to detect the complete features of peripheral targets because it overemphasizes the center, while our algorithm makes target detection complete via the inclusion of local energy tunnel to emphasize target outlines and weaken the property of Gabor function to highlight the center. Similarly, the result shown in the fourth row also shows the advantages of our algorithm. The image is a multiple-target scene image. The result shows that Itti Algorithm throws a few false positive detections and, compared to our algorithm, the necks of giraffes on the sides are confused with the background.

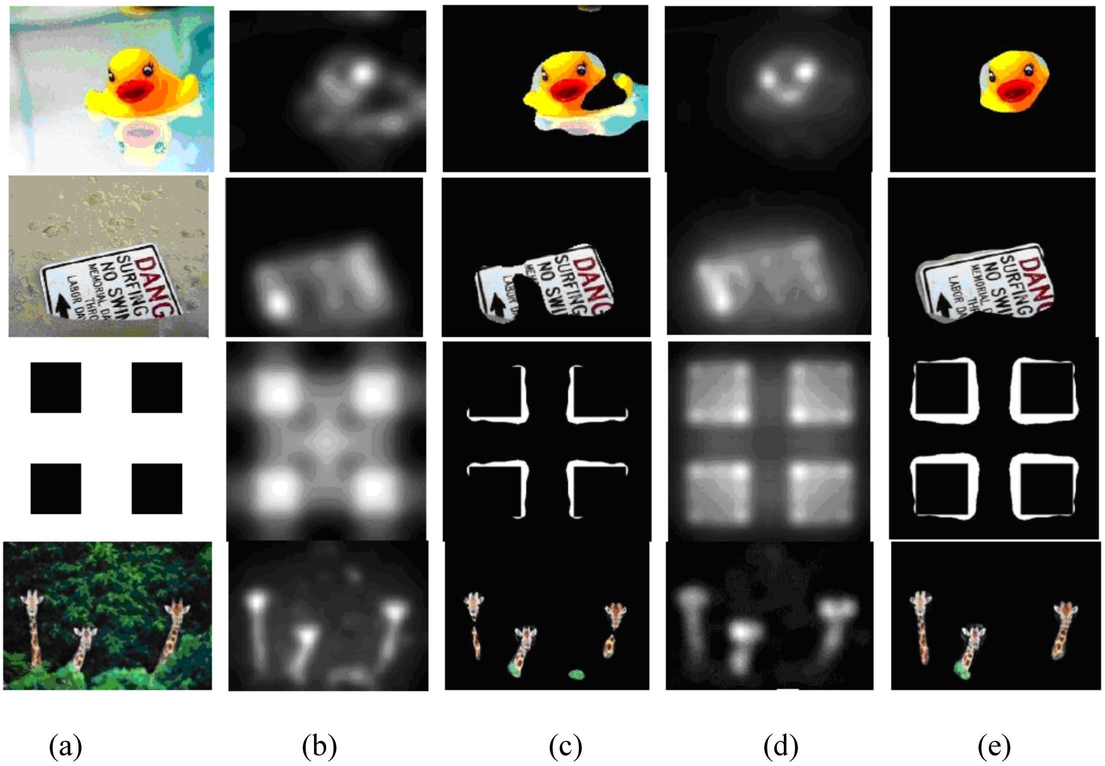


Figure 4 Comparison of experimental results with different methods((a) Originalimages; (b) Saliency maps of Itti' s method; (c) Salient objects of Itti's method;(d) Saliency maps of the proposed method;(e) Salient objects of theproposed method)

Next, we use objective methods to test the salient target detection method in this article. As shown in Figure 4, the discovery rate and accurate rate in our algorithm are both higher than those in Itti model. We employ F-measure for better judgment. Its formula is:

$$F\_measure=\frac{(1+\alpha)*\Pr ecision*\operatorname{Re}call}{\alpha*\Pr ecision+\operatorname{Re}call} \tag{23}$$

In this formula, *Precision* is the accurate rate, *Recall* is the discovery rate and α=0.3. It is shown in Figure 5 that the F-measure of our method is higher than Itti model.

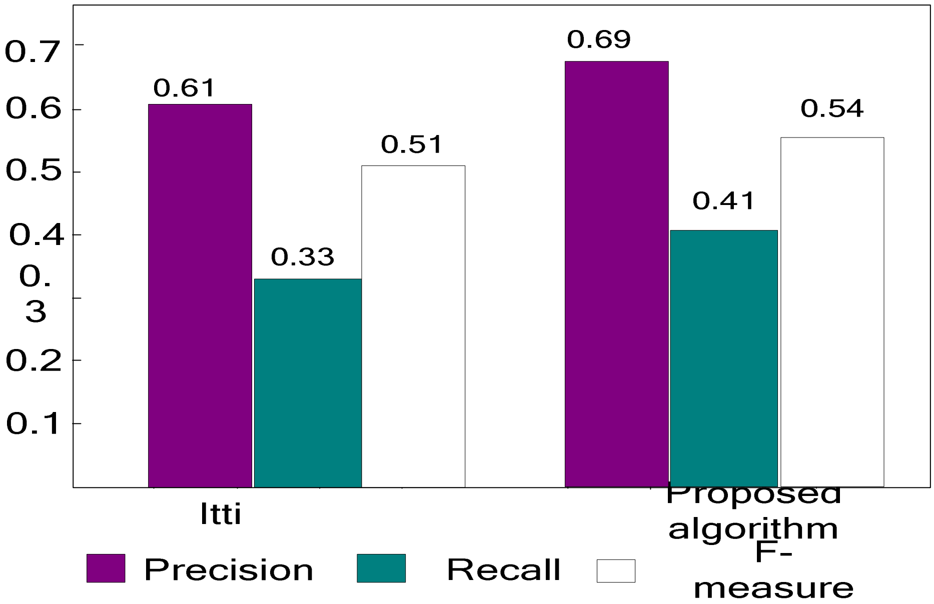


Figure 5 the discovery rate, accurate rate and F-measure of Itti Model and our algorithm

B. Experimental comparisons with other target detection methods

Experiment 1: Compare the improved saliency map method with Saliency Tool Box (STB), Spectral Residual (SR), Frequency-tuned (FT) and Context-aware Visual Saliency, and the result is shown in Figure 6. Evaluating subjectively, the saliency map constructed via our method is obvious better than STB. SR method is better than our method on the first three images whose backgrounds are relatively simple, but our method is better when the background is noisy, such as in the giraffe image. The saliency maps of FT are not as good as our method in all four images. Context-aware Visual Saliency has ineffective performances in the giraffe image and the square image. Objective evaluation according to AUC is given in Table 1, which shows that the saliency map of our method is better than other models. From the Table 1, our proposed algorithm is excellent than other previous algorithm.

Table1 AUC of Different Models

Model	STB	SR	FT	CA	Our proposed
ACU	0.732	0.786	0.746	0.779	0.807

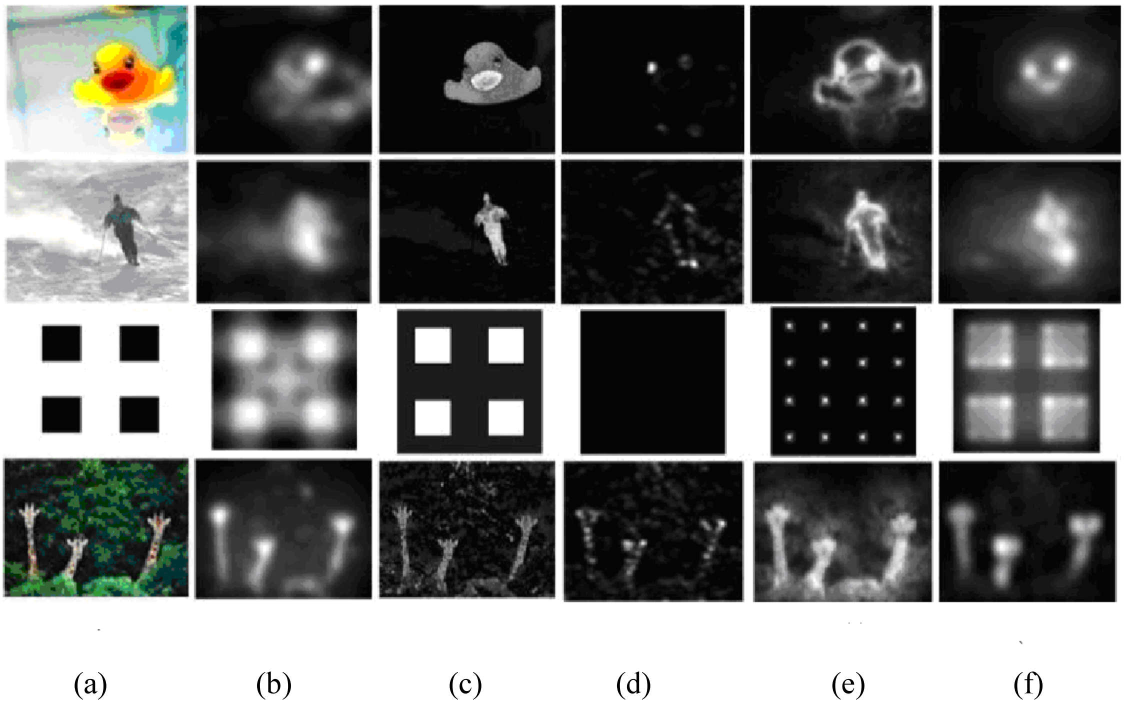


Figure 6 Comparison of saliency maps

Experiment 2: images of successive frames are used as experimental samples, which are from Kazuma Akamine. We compare our algorithm with Kazuma Akamine and Itti. To improve the algorithm, we have included an optimized factor:

$$P_{est} = 4k(M_1)^{1/2} / M \tag{24}$$

The above is the optimized factor for saliency detection, which mainly helps to refine the salient target region. k is an adjustable parameter. Suppose the salient target region is a square with M_1 pixels, then the side length of this region is about $4(M_1)^{1/2}$. M is the number of pixels in the saliency map and P_{est} is the estimation of the number of pixels around the salient target region.

Table 2 Comparison on target detection

Method		Precision	Recall	F-measure
Itti		0.784	0.821	0.802
K.Akamine		0.849	0.887	0.864
Our method	K=2	0.879	0.886	0.874
	K=3	0.873	0.908	0.892
	K=5	0.896	0.857	0.872

V. CONCLUSIONS

In this article, we combined visual saliency model with image segmentation and proposed an adaptive salient target segmentation method to realize the separation of salient targets from complicated backgrounds. In the visual saliency model, the addition of local energy to describe the outline of targets maintains the completeness of salient targets. During the adaptive segmentation, we used Maximum Entropy Criteria to estimate the target entropy, then we calculated the entropies of pre-segmented images constructed from saliency maps, and finally we selected the optimal target detection result according to the entropies. On the basis of visual attention model and physiology, we included local energy to the saliency model to describe the outlines of targets, and then we separate targets and backgrounds according saliency and estimate the salient target region via Maximum Entropy Criteria. A number of experiments have proved that our algorithm provides more complete salient target detection and higher discovery rate, accuracy rate as well as F-measure than tradition methods. This method is applicable to visual salient target detection in complex backgrounds.

REFERENCES

[1] Xingting Gao, Sattar F, and Venkateswarlu R. Multiscale corner detection of gray level images based on log-Gabor wavelet transform, Circuits and Systems for Video Technology, IEEE Transactions on, vol.17, pp. 868-875, 2007.

[2] He X C, Yung N H C. Curvature scale space corner detector with adaptive threshold and dynamic region of support[C]//Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. IEEE, pp. 791-794, 2004.

[3] A Nachar, R.; Inaty, et al.,A robust edge based corner detector, 2014 17th IEEE Mediterranean Electro-technical Conference (MELECON), pp.242-246, 2014.

[4] Saxena A, Sun M and Ng A Y. Make 3d: Learning 3d scene structure from a single still image, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 31, pp. 824-840, 2009.

[5] Zhu Feng, Feng Yiping, et al., Information integration strategy of the petrochemical industry from the multi-scale perspective, 2013 10th IEEE International Conference on Control and Automation (ICCA), pp.1284-1289, 2013.

[6] He C., et al., Local Topographic Shape Patterns for Texture Description, IEEE Signal Processing Letters, vol. 22, no.7, pp. 871 - 875, 2015.

[7] Lowe D.G. Distinctive image features from scale-invariant key points, Proceedings of International journal of computer vision, 2004, 60(2): 91-110.

[8] Kisku, Dakshina Ranjan et al., Face image abstraction by Ford-Fulkerson algorithm and invariant feature descriptor for human identification, 2014 International Carnahan Conference on Security Technology (ICCST), pp.1-4,2014..

[9] Andrea Perna, Michela Tosetti, Domenico Montanaro, et al. BOLD response to spatial phase congruency in human brain, Vision of Journal, 2008, 8(10):1-15.

[10] Linda Henriksson, Aapo Hyvarinen, Simo Vanni. Representation of Cross-Frequency Spatial Phase Relationships in Human Visual Cortex, the Journal of Neuroscience, 2009, 29(45):14342-14351.

[11] S Dustin, E. Stansbury, Thomas Naselaris, Jack L. Gallant, Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex, Neuron, vol. 79, no. 5, pp. 1025–1034, 2013.

[12] Michael Felsberg, Gerald Sommer, The monogenic signal, IEEE Transactions on Signal Processing, vol. 49, pp. 3136-3144, 2011.

[13] N. Tamayo, V.J. Traver, Entropy-based Saliency Computation in Log-polar Images, 3rd International Conference on Computer Vision Theory and Applications, Madeira, Portugal, pp. 501-506, 2008.

[14] G. Backer, B. Mertsching. Two selection stages provide efficient object based intentional control for dynamic vision, International Workshop on Attention and Performance in Computer Vision, pp. 9-16, 2003,

[15] Wei Wei, Shen Xuanjing, Qianqingji.et al. Thresholding algorithm based on three-dimensional Renyi's entropy, Journal of Jilin University (Engineering and Technology Edition), 41(4):1083-1088, 2011.

[16] Mingming Cheng, Guoxin Zhang, Niloy J. Mitra, et al. Global Contrast based Salient Region Detection , IEEE CVPR, Colorado Springs, vol. 21-23, pp. 409-416, 2011.

[17] Huanbing Gao, Shouyin Lu, Guohui Tian, Jindong Tan, Vision-integrated physiotherapy service robot using cooperating two arms, International Journal on Smart Sensing and Intelligent Systems, vol.7, no.3, pp.1024 – 1043, 2014.

[18] Archana S. Ghotkar and Dr. Gajanan K. Kharate, Study of vision based hand gesture recognition using Indian sign language, International Journal on Smart Sensing and Intelligent Systems, vol.7, no.1, pp. 96 – 115, 2014.