Robot Visual Tactile Information Fusion Drives Visual Attention

Jian-gang SUN¹, Tao YU², Chun-chun MA¹ and Hui WANG¹

¹Beijing Mining and Metallurgical Technology Group Co., Ltd., Beijing, China
²Beijing Aerospace Changfeng Co., Ltd., Beijing, China

Keywords: Visual attention, Attention focus, Saliency, Visual touch fusion.

Abstract. This paper describes a theoretical model of visual tactile information fusion mode to drive visual attention. According to the visual and tactile information obtained by the robot, the corresponding position of the most important object and the stimulation signal point for the current task is determined, and the visual attention is focused on the target object of interest. According to the error information fed back by the sensor, the relevant parameters are continuously corrected, and the robustness of the system is improved. The theory was tested on a robot with two degrees of freedom on the head and two degrees of freedom on the arm. The test results show that the model can adapt to changes in itself and the environment and has good robustness.

Introduction

 Humans receive a large amount of information from the visual, tactile, auditory, and olfactory at every moment in the course of their homework or behavior. They drive the head by comparing the saliency of the obtained information so that the region of interest is located. The high-resolution foveal area of the retina allows attention to the area so that accurate visual perception of the area can be achieved. This is called selective attention in cognitive physiology. Studies have shown that humans and many animals have mechanisms that can fuse multimodal information to stimulate visually selective attention and quickly retrieve regions of interest. Regions of interest are the areas that are most interesting and best represent the content of the image [1-3]. The precise visual perception of the region is to obtain local features of the image, eliminate interference from secondary information, and obtain more accurate image content.

An intelligent robot with a multi-mode sensing system can collect a large amount of sensor information, which lays the foundation for the robot to issue the correct instructions. However, processing a large amount of information reduces its response speed. How to improve the response speed of robots based on making full use of sensor information resources has become a hot topic in robot research. Imitating human multi-modal information fusion to drive visual attention mechanism, using robot multi-modal information fusion to drive visual attention is an effective way to solve this problem. It can use sensor resources more fully and reasonably, and take advantage of visual information to make the information obtained by visual sensors more accurate and more valuable. Avoid processing a large amount of information unrelated to the current task, and improve the efficiency of information processing. In order to adapt the robot system to the changes of the surrounding optical environment and the wear of the mechanical system, simulate the process of infant intelligence development, use the feedback sensor information, correct relevant parameters in real time, and compensate for errors caused by changes in the external and own conditions.

Selective attention is one of the hot spots in the field of machine vision research. At present, there are a large number of related literatures on visual information-driven visual attention [4-6], using the color, brightness, texture and other related information collected by the visual sensor to calculate the corresponding saliency and drive attention shift. According to the characteristics of the salient points, the region of interest close to the salient point features is retrieved, and the region of interest is image processed, which reduces the area of image processing and improves the efficiency of computer image information processing. In recent years, neurophysiological research has obtained some research results on multi-modal information fusion to drive visual attention. In particular, the latest research results in brain science indicate that there are regions of multi-modal
information fusion inside the brain [7,8], which inspired Robot vision and tactile information fusion drive the study of visual attention.

**Inspiration from Neurophysiology**

Humans have a multi-mode selection attention mechanism, which is significant for inspiring us to study the meaning of robot vision and tactile information fusion to drive visual attention. Every day in our lives, multi-mode perception events occur: when we see someone else’s mouth moving, they will hear their voices; when our fingers are pierced by the needle tip and feel pain, the eyes will see puncture. The location oozes blood. In the 1990s, brain researchers discovered that there is a single pattern of sensory information processing within the brain, that is, areas that deal with visual, tactile, auditory and other perceptual information [9]. In recent years, brain science researchers have further discovered that there are also areas for processing multi-modal information fusion inside the brain, especially for the intersection of visual, auditory and tactile patterns. The Beauchamp MS [7] researcher's hypothalamus discovery, in addition to the individual information processing areas of perception modes such as vision, touch and hearing, there are also areas of visual and tactile information fusion processing; visual and auditory information fusion processing areas; visual, auditory, and tactile information fusion processing areas.

Vision and touch are the primary modes of perception of humans and other animals, producing a consistent description of objects. As shown in Figure 1, under normal circumstances, when a person's finger is burned by a very hot cup, he immediately focuses the visual attention on the glass and extracts its distinctive features. When he sees the glass again next time, he will subconsciously feel the pain of his hand; at the same time, when he is burnt again, the image of the cup will appear immediately in the brain. As the saying goes: “I was bitten by a snake, and I was afraid of a well rope for ten years.” In a certain area of the brain, visual and tactile perception information can be uniformly encoded to process the fusion information obtained by various perceptual modes.

Psychologists have done a lot of experiments and discovered the theory of multimodal information fusion for selective attention. For example, Spence et al found that hearing can cause visual attention and drive the shift of visual attention [10]; Eimer further improved the related theories of visual and tactile cross-mode driving visual attention based on the work of predecessors [11]. According to the research results of psychology, Spence [12] proposed the theory that visual and tactile fusion modes drive visual attention. When both visual and tactile perception signals are obtained, the most interesting target object is judged in the brain based on the collected information. Drives the movement of the head, moving the highest resolution retinal fovea to the centroid of the object of interest, accurately sensing the visual properties of the object. Related experiments with animals such as monkeys and cats have shown that other animals also have mechanisms similar to human multimodal perception information driving visual attention.

However, people are not born with a sophisticated multi-modal information fusion mechanism that drives visual attention. When a baby is born, he can't judge the position of the visual and tactile stimulation points well. As he grows up and interacts with the environment, he gradually clears the
approximate location of the visual stimulation points and tactile stimulation points, and focuses on the visual attention. Focus on the stimulus target of interest. For example, when a baby touches a doll with his hand, he slowly shifts the focus of visual attention to the doll. However, during his long period of time, he could not accurately complete the tactile-driven visual attention and make accurate visual perception. Within his visual perception, he also cannot accurately focus the attention of his attention on the area he is interested in. Only after a long period of continuous learning and training can the visual tactile information fusion be well driven to drive visual attention. The process of infant growth is the process of infant intelligence development. The learning mechanism of infants provides important clues for us to study the complex process of machine learning. According to theories of sensorimotor development proposed by developmental psychologists and robotics researchers, the learning ability of infants shows that a combination of kinematics, dynamics and neurophysiology can make a developmental body and the outside world The random interaction interacts to optimize the visual and tactile information fusion with learning ability to drive the visual attention mechanism.

Inspired by the visual attention mechanism stimulated by human visual and tactile fusion patterns and the development of infant intelligence, this paper establishes a theoretical model of visual tactile information fusion driving visual attention with learning mechanism. The robot was simulated with a test system with two degrees of freedom of the head and two degrees of freedom arms, and related tests were carried out. The experimental results show that the model is feasible and has good robustness.

**Visual/Touch Dual Mode Driven Visual Attention Model**

In order to be easy to understand, this paper first gives a schematic diagram of visual and tactile information fusion driving visual attention. As shown in FIG. 2, the perception system of the robot is mainly composed of a visual perception system and a tactile perception system. The head with the vision sensor has two degrees of freedom, namely the angle of rotation $\beta$ around the neck joint and the pitch angle $\gamma$. The arm with the tactile sensor also has two degrees of freedom, $\delta$ around the shoulder joint and $\theta$ around the elbow joint. As shown in the figure, the visual sensor can sense the color, shape, size, brightness, orientation and the like of the object; at the same time, the tactile sensor can sense the temperature, hardness, shape, roughness and the like of the object.

![Figure 2. Visual haptic information fusion driven visual attention diagram.](image)

According to the current task, the various external information obtained is different to the robot, and the difference is expressed by saliency. The level of significance indicates that the corresponding attribute of the object is more important to the current task. Define the saliency of each kind of perceptual information, calculate the comprehensive saliency of each object, determine the focus of visual attention, and use a sensor's information to drive visual attention, as shown in Figure 2, in the integrated contrast robot. After the information about the shape, color, hardness, temperature, etc. of the object, according to the task, the red ball is determined as the focus of attention, and the position of the centroid of the object is obtained by the visual sensor, so that the optical axis of the camera coincides with the centroid of the ball. The range observed by the vision
sensor is the area of interest (ROI) of the robot. When performing accurate visual perception, only the area of interest is processed, which reduces the task of image processing and greatly improves the speed of image processing. At the same time, the local salient features of the object can be accurately obtained.

Vision is the most important perception mode of robots. When the most prominent object has two sensing modes to obtain its distinctive features at the same time, the visual perception mode is selected to drive visual attention. The visual image can obtain the position information of the object centroid. Through visual calibration, the actual positional relationship between the camera and the object can be obtained, and the robot head movement can be driven to make the optical axis of the camera coincide with the position of the object centroid, and the visual camera is High-resolution regions are concentrated in regions of interest with similar features. When the object with the highest degree of sensibility only perceives it, it drives the visual attention through the tactile information. Through the coordinate transformation, the positional relationship between the object and the camera is obtained, and the robot head is driven to make the optical axis of the camera coincide with the centroid of the object.

When the robot is working in a completely new working environment, the lighting environment changes and the accuracy of driving visual attention is affected. At the same time, due to the errors caused by calibration errors and mechanical system wear, the motion accuracy of the robot head will also be affected. Simulating the process of infant intelligence learning, using machine self-learning theory, according to the camera to obtain error information generated by each head movement, respectively, to compensate for the error caused by the optical system and mechanical system. The system can continuously improve the accuracy of visual attention through a period of learning and training under the condition of simple visual calibration and simple measurement of the dimensions of the mechanical components.

Choose to Pay Attention to the Establishment of Mathematical Models

Inspired by the latest research results of neurophysiology, according to the visual and tactile crossover mode established by M.Rucci [13] to drive the theoretical model of visual attention, this paper establishes a theoretical model of visual and tactile information fusion to drive visual attention, as shown in Figure 3.

M.Rucci's theoretical model implements visual attention driven by visual, tactile crossover mode. That is, according to the current task, compare the visual perception of each object with the saliency of the tactile perception information, and then compare the maximum saliency values of all objects in the saliency map, and select objects with greater saliency in all objects to drive visual attention. The model established in this paper is to compare the saliency of visual and tactile information fusion, and select the object with greater significance as the object that drives visual attention. The region with the characteristic feature close to the object is regarded as the region of visual interest, and the optical axis of the camera coincides with the centroid of the object, and the image of the region is processed. The processing of image information unrelated to the current task is avoided, the calculation amount of information processing is reduced, and more accurate image information is obtained. At the same time, according to the error information returned by feedback, the relevant parameters are continuously corrected, and the robustness of the system is improved.
The camera position map encodes the position of the current camera. The visual information map encodes the original information obtained by the camera, and extracts its corresponding salient features to obtain the saliency of all objects within the visual perception range in the visual saliency map. The tactile information map encodes the original information obtained by the tactile sensor, and extracts its corresponding salient features to obtain the saliency of all objects within the perceptual range of the tactile sensor in the tactile saliency map. The saliency of each object is calculated on the saliency map, that is, the combination of visual saliency and tactile saliency. The object with the most significance is determined as the target object that drives the camera motion, and the position of the object is obtained in the position map of the object to determine the correspondence between the current camera position and the object position. The head movement is driven such that the optical axis of the camera coincides with the center of the region of interest. After completing the motion of driving the camera once, the camera obtains the motion error, corrects the corresponding parameters, and gradually compensates for the motion error.

The visual map encodes the original image obtained by the camera. The degree of significance of information such as color, grayscale, orientation, shape, and the like of the image is obtained in the visual saliency map. The value of the saliency is set in advance according to the current task. As shown in Table 1. Each object causes a significant degree of visual attention due to visual information. which is:

\[ \alpha = \sum_{i=1}^{n} a_i \]  

Table 1. Significant degree of visual information.

<table>
<thead>
<tr>
<th>Image information</th>
<th>colour</th>
<th>Grayscale</th>
<th>shape</th>
<th>position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significant degree</td>
<td>(a_1)</td>
<td>(a_2)</td>
<td>(a_3)</td>
<td>(a_4)</td>
</tr>
</tbody>
</table>

The tactile information map encodes the tactile signal, and the original information related to the temperature, hardness, roughness, humidity, and the like of the object can be obtained. The salient features obtained by the tactile sensor are extracted, and the tactile saliency map obtains the saliency of the tactile corresponding information. The value of the saliency is preset in accordance with the importance of the corresponding haptic information in the current task. As shown in table 2. Thereby, the degree of saliency of each object due to tactile perception is obtained. which is:

\[ b = \sum_{j=1}^{n} b_j \]  

Table 2. Significance of tactile information.

<table>
<thead>
<tr>
<th>Tactile information</th>
<th>temperature</th>
<th>hardness</th>
<th>Roughness</th>
<th>humidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significant degree</td>
<td>(b_1)</td>
<td>(b_2)</td>
<td>(b_3)</td>
<td>(b_4)</td>
</tr>
</tbody>
</table>

Calculate the saliency of each object under the current task based on the current task of the robot. which is:

\[ S = \alpha + b \]
Comprehensively compare the saliency S of each object under the current task. Select a significantly larger object as the target object that causes visual attention, and define a region with significant features as the region of interest. Selecting a certain mode of perception of the object drives visual attention, driving the head movement with the camera, so that the optical axis of the camera coincides with the center of the region of interest, enabling precise visual perception of the region.

Driving visual attention mode 1: Vision as a mode to drive visual attention. The current position of the camera is obtained by a motor encoder that drives the movement of the head. By processing the image, the pixel coordinates of the target object are obtained. Through calibration, the positional relationship between the target object and the camera is determined. That is, the pixel reference coordinates of the object are converted into position coordinates centered on the optical axis of the camera. That is, the corresponding transformation is:

\[ V^x(t) = F_r(V_{ij})w^xK_xM + K_yM(k_1M^2 + k_2M^4) \]  
\[ V^y(t) = F_r(V_{ij})w^yK_yM + K_xM(k_1M^2 + k_2M^4) \]

The step function is the termination condition for driving the camera motion. When the distance between the object centroid and the optical axis of the camera reaches a preset value, the motion of driving the camera will be terminated. \( w \) is the camera calibration coefficient, and the calibration coefficient can be set according to experience. \( K_x\) with \( K_y \) is the component coefficient is the distance between the object centroid and the optical axis of the camera in two degrees of freedom; \( M \) is the pixel distance between the optical axis of the camera and the object centroid. \( k_1 \) with \( k_2 \) To correct the coefficient of radial distortion, radial distortion is the main form of image distortion. The correspondence between the center of the region of interest and the position of the camera is obtained by the above formula.

Due to errors in mechanical systems, optical systems, etc., it is impossible to obtain accurate driving visual attention related parameter information by calibration. When the environmental information such as light changes greatly, it will further affect the system to obtain more comprehensive and accurate visual information of the object. Correct weights by learning in real time. Can further improve the robustness of the system. The driving motion of visual attention is completed, and the motion error obtained by the feedback camera is used to compensate the error caused by the mechanical system and the optical system in real time to adapt to changes in the environment and the like. The weight of the system is corrected by (6) and (7):

\[ w^x(t + 1) = w^x(t) + k^x_v \epsilon_x w^x(t)/V_{ij}^x(t) \]  
\[ w^y(t + 1) = w^y(t) + k^y_v \epsilon_y w^y(t)/V_{ij}^y(t) \]

where \( k \) is the learning coefficient between (0-1), and the learning process is as shown in equations (8) and (9). \( \epsilon_x \) with \( \epsilon_y \) It is the error produced in the two directions of motion.

\[ k_1(t + 1) = k_1(t) + k^x_v(\epsilon_x/(x_1M^2_1 - 1/x_1x_2M_1^4 + x_1x_2M_2^4)) \]  
\[ k_2(t + 1) = k_2(t) + k^y_v(\epsilon_y/(x_2M^2_2 - x_1M_1^2 + x_1x_2M_2^4)) \]

Driving visual attention mode 2: Tactile as a mode to drive visual attention. The position information of the sensor is obtained by the servo motor encoder, thereby obtaining the position coordinates of the target object touched by the sensor, and the positional relationship between the target object and the camera is obtained by coordinate transformation. The change relationship is:

\[ T^x(t) = F_r(T)(M^x + Z^x(t)) \]  
\[ T^y(t) = F_r(T)(M^y + Z^y(t)) \]

\( F_r(T) \) It is a pre-valued function. When the distance between the centroid of the object and the optical axis of the camera reaches a preset value, the motion of driving the camera will be terminated. Extract local features of the region of interest. \( M^e \) is the tactile sensor position coordinate, \( Z(t) \) is the geometric parameter of the relevant mechanical component.
Through the coordinate transformation of (10) and (11), the position coordinates of the object in
the pixel reference coordinate system of the camera are obtained, and the visual attention is driven
to perform accurate visual perception.

Due to errors in mechanical system fabrication and measurement, there is an error in the motion
of the robot to drive visual attention from the tactile perception mode. After completing the driving
motion of visual attention, the motion error is obtained through the analysis of the image, and the
relevant parameters are corrected in real time by the motion error of the feedback to compensate for
the motion error:

\[ Z^x(t + 1) = Z^x(t) + k_1 \delta^x \]

\[ Z^y(t + 1) = Z^y(t) + k_2 \delta^y \]

where \( k \) is the learning coefficient, \( \delta^x \) with \( \delta^y \). The tactile information drives the camera to produce an
error in the direction of motion.

**Experiment and Analysis of Visual Tactile Information Fusion Drives Visual Attention**

In order to verify the effectiveness of the visual and tactile information fusion driving visual
attention theory, we establish a test platform. The basic structure of its hardware is shown in Figure
5: Using the above X-axis and Y-axis to simulate the head of the robot, the CCD camera can be
moved in the X-axis and Y-axis directions. The X and Y axes below simulate the robot's arm and
can drive the tactile sensor along the X and Y axes. We used the test rig shown in Figure 6 to
perform visual and tactile information fusion to drive visual attention testing. By comparing the
saliency of each object, the object with the highest degree of significance is selected as the object
that causes visual attention, and a perception of the object is selected to drive visual attention. The
camera is driven to move so that the centroid of the target object of interest is close to the optical
axis of the highest resolution camera. The effect of selecting visual mode and tactile mode to drive
visual attention was verified separately. After several trainings, the accuracy of driving the camera
movement is gradually improved. The optical axis of the camera is approximated to the center of
the region of interest, and image processing is performed on the region of interest, which avoids the
processing of current task-independent information and reduces the task of image processing.
Because vision is the main mode of robot perception. Therefore, the degree of significance of visual information is high, and the degree of significance of tactile information is low. The distribution of its significance is shown in Table 3 below. When the visual camera perceives that the redness is 0.6, the tactile sensor senses the metal object with a degree of significance of 0.5. The visual and tactile sensors perceive that other information has a significance of zero.

Table 3. Setting of saliency.

<table>
<thead>
<tr>
<th>Perceived information</th>
<th>red</th>
<th>metal</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significant degree</td>
<td>0.6</td>
<td>0.5</td>
<td>0</td>
</tr>
</tbody>
</table>

In order to improve the speed of driving visual attention, we use a small amount of image processing methods. We preliminarily specify the color of the object of interest, apply color clustering to simplify the color information, and perform approximate color matching. Figure 7 shows the original image taken by the camera. By using this method, the RGB color space is divided into several color subspaces. Here we summarize a color space consisting of 25 colors in Table 1. They are extracted from the standard samples of RGB three primary colors [14]. Find the values of the three primary colors that are similar to the colors that are preset to cause visual attention points in Table 1. In Table 1, find the values of the three primary colors that are similar to the colors that are set to cause visual attention points, and use them separately. $C_{iR}, C_{iG}, C_{iB}$ said to use $P_r, P_g, P_b$ indicates the values of the three primary colors of red, green, and blue in the image. $C_d$ is calculated by formula (14)

$$
C_d = \sqrt{(P_r - C_{iR})^2 + (P_g - C_{iG})^2 + (P_b - C_{iB})^2} \quad (14)
$$

Table 4. Standard composition of the extracted ternary colors of 25 colors.

<table>
<thead>
<tr>
<th>Color</th>
<th>R</th>
<th>G</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sea green</td>
<td>0</td>
<td>182</td>
<td>0</td>
</tr>
<tr>
<td>Light green</td>
<td>0</td>
<td>255</td>
<td>170</td>
</tr>
<tr>
<td>Olive green</td>
<td>36</td>
<td>73</td>
<td>0</td>
</tr>
<tr>
<td>Aqua</td>
<td>36</td>
<td>146</td>
<td>170</td>
</tr>
<tr>
<td>Bright green</td>
<td>36</td>
<td>255</td>
<td>0</td>
</tr>
<tr>
<td>Blue</td>
<td>73</td>
<td>36</td>
<td>170</td>
</tr>
<tr>
<td>Green</td>
<td>73</td>
<td>146</td>
<td>0</td>
</tr>
<tr>
<td>Turquoise</td>
<td>73</td>
<td>219</td>
<td>170</td>
</tr>
<tr>
<td>Brown</td>
<td>109</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>Blue gray</td>
<td>109</td>
<td>109</td>
<td>170</td>
</tr>
<tr>
<td>Lime</td>
<td>109</td>
<td>219</td>
<td>0</td>
</tr>
</tbody>
</table>
The filtering rules are defined as follows:

$$
\begin{align*}
C_0 &= C_r, & C_d \leq C_i \\
C_0 &= \text{Black}(RGB(0,0,0)), & C_i > C_r
\end{align*}
$$

(15)

$C_0$ is the output of the filtered pixel, $C_r$ is the color information that causes visual attention, $C_i$ is the set color distance. After filtering, the image is binarized. The test results are shown in Figure 3:

![Figure 8. Image processed by assigning weights and images.](image1)

![Figure 9. Image of the attention result.](image2)

Calculate the saliency of each object to determine the target object that causes visual attention. Selecting a mode for its perception drives visual attention and obtains salient features of the object of interest.

Use vision as a model to drive visual attention. The position coordinates of the camera can be obtained by the motor encoder. After image processing, the position coordinates of the centroid of the object are obtained, thereby obtaining the pixel distance between the optical axis of the camera (the center of the image) and the centroid of the object, and after calibration, it is converted into the parameter of the encoder that drives the camera motion. Achieve visually driven visual attention.

The effect diagram of driving visual attention is shown in Figure 4, and the center of the object is closer to the optical axis of the camera.

At the beginning of the training, due to the mechanical system, optical system and calibration errors, the motion accuracy of driving visual attention is low. After a period of training, the visual calibration parameters gradually become constant, and the visual attention error is also reduced to a certain range. As shown in Figure 5 and Figure 6, the calibration coefficients tend to be constant. The error is reduced from 0.32619 to 0.0031. When the mechanical system and the light system change, after training, the expected visual attention will still be achieved.
Tactile sense is used as a mode to drive visual attention. The position of the tactile sensor, i.e., the position of the object coordinates, can be obtained by the servo motor encoder of the arm. Using the theory of robotics, spatial coordinate transformation is performed, and the coordinate position of the object sensed by the tactile sensor is converted into the coordinate under the reference coordinate system with the camera as the coordinate point, and the obtained coordinate value is converted into a servo motor that drives the camera motion. Set parameters to drive camera motion. Thereby, the optical axis of the camera is coincident with the position of interest of the tactile sensor.

Due to the errors and motion errors of the coordinate transformation, the accuracy of the haptic-driven visual attention is low at the beginning of the training. After a period of training, the system gradually reduces the error and enables more accurate visual perception. As shown in Figure 7. This shows that the system can adapt to changes in the environment, compensating for errors caused by system and coordinate transformation.

From the above image of the visual calibration coefficient and the error image of the visual and tactile driving visual attention, it can be seen that as the number of trials increases, the visual calibration coefficient gradually becomes within a reasonable numerical range. At the same time, the visual haptics drive the visual attention error gradually decreasing.

In experiments in which visual tactile fusion drives visual attention, when a tactile sensor perceives a series of object tactile information, in order to obtain more information that drives visual attention, further visual information needs to be obtained. According to the position coordinates of the object provided by the tactile sensor, when the image information is obtained by the camera, it is only necessary to perform image processing on the region where the related object is located. As shown in Fig. 8, in order to detect a red metal object, if only visual perception is required, all the information of (a) map needs to be processed, the processing time is about 165 ms, and when the position information of the object according to the tactile feedback is processed, only
the map is processed (b) In a few small areas, the processing time is only 8ms. The speed of processing images is increased by 20 times.

![Image](image.png)

Figure 13. Image processing for tactilely driven visual attention.

## Conclusion

This paper simulates the multimodal information fusion between humans and other animals to drive the mechanism of visual attention and the process of infant intelligence development. A theoretical system of multi-modal information fusion with learning mechanism to drive visual attention has been established. The architecture integrates visual and tactile mode information, drives visual attention according to the current tasks of the system, and performs precise visual perception. It can obtain more accurate and prominent features of objects that cause visual attention, and provide more detailed visual information for the current task of the robot, avoiding other interferences that are not related to the current task improve image processing speed. Through the image processing, the error of the camera motion can be obtained. Through the error value of the feedback, the error caused by the changes of the environment and the system can be compensated in real time, and the system is more intelligent and more robust. The verification test of the theory was carried out on the test platform, and the ideal test results were obtained.

Neurophysiological research on the multi-modal information fusion of human and other animals drives the visual attention mechanism and the development of infant intelligence, laying the foundation for the study of multi-modal information fusion of intelligent robots with learning mechanisms. According to the latest research results of neurophysiology, the visual tactile information fusion model and visual learning model are driven by the fusion of visual tactile information, but they are all based on simpler conditions and still need further improvement. The author believes that there are several possible research directions in the future of this research field: First, in the face of complex task visual tactile information fusion driven visual attention theory research, further improve the system's effectiveness and robustness. Second, the research on the visual attention mechanism driven by the visual-acceptance fusion mode with learning mechanism.

## References


