Task-driven Saliency Detection on Music Video

Shunsuke Numano, Naoko Enami, Yasuo Ariki

Kobe University

Abstract. We propose a saliency model to estimate the task-driven eye-movement. Human eye movement patterns is affected by observer's task and mental state[1]. However, the existing saliency model are detected from the low-level image features such as bright regions, edges, colors, etc. In this paper, the tasks (e.g., evaluation of a piano performance) are given to the observer who is watching the music videos. Unlike existing visual-based methods, we use musical score features and image features to detect a saliency. We show that our saliency model outperforms existing models that use eye movement patterns.

1 Introduction

Yarbus suggested that human eye-movement patterns are modulated top down by different task demands[1]. After that, many works had showed the relationship between eye-movement patterns and cognitive factor[2][3][4]. From these analyses, Itti *et al.*[4] proposed two hypotheses which are called the saliency hypothesis with relation to eye-movement. Hypothesis 1 is that the eye-movement patterns is affected by the task such as driving. Hypothesis 2 is that human gazes at the area which has the saliency of color, edge and the intensity when they are not given the task. The method for the estimation of the task such as the documents in reading [5] using eye-movement patterns is proposed based on hypothesis 1. However, no method is proposed that estimates the eye-movement patterns for different tasks from the image . On the other hand, based on hypothesis 2, Itti *et al.*[6] proposed the method for the estimation of the "saliency" area where human gazed in terms of the image feature such as color, intensity and orientation. However, the existing saliency model are not considered the hypothesis 1.

Our goal is to detect the image saliency so that we can estimate the task-driven eye-movement. In this work, we give the observers two tasks when viewing the music videos. Task 1 is to evaluate the piano performance, and task 2 is to grasp the melody of the performance. These tasks are affected by the cognitive factor related to music. We can introduce the cognitive factor to the evaluation of the musical sense that can be obtained in the musical education. Thus, we need to show the relationship between the cognitive factor and the eye-movement patterns related to music. In accordance with above, we should consider our tasks as well as the conventional image feature when constructing the saliency model. In addition, the observers listen to the sound as well as viewing the image in

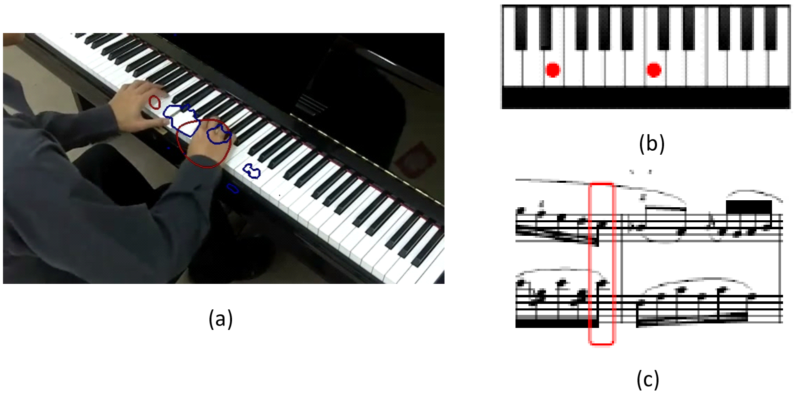

Fig. 1. Our saliency model adding the musical information. (a) Our saliency model(blue line), the ground truth(red line). (b) The striking key on the keyboard. (c) The striking key on the musical score (red box).

our tasks. So, we expect that the eye-movement is affected by the music video including the sound and the tasks.

Although most of existing saliency models are detected from the low-level image features, the saliency affects the observer's knowledge (e.g., the musical sense in this work) to the observation object and the cognitive factor. In this paper, we propose a novel saliency model that is added the information of the musical score in order to achieve our goal as shown in Fig.1(a). The musical score has a lot of information necessary for the performance such as the musical note, the dynamics, and so on, and the performance is conducted in accordance with the musical score. We therefore can add more information related to the performance by using the musical score information than the sound information, and we expect to obtain the saliency related to the knowledge of the music. In this paper, we add the information of the musical note (as shown in Fig.1(b),(c)) to the conventional saliency map that is proposed by Itti *et al.*[6] who detected the saliency using the image feature and that is the state-of-the-art proposed by Yang *et al.*[7]. Next, we evaluate the proposed saliency model. Our goal is to construct the saliency model for the estimation of the task-driven eye-movement, so we use the task-driven eye-movement as the ground truth in the evaluation. The method where the eye-movement is used for the estimation of the saliency map that is detected from the video is proposed [8]. However, Riche *et al.*[8] used the eye-movement that was not considered the task. Thus, considering the task in this paper, we construct the dataset which is constituted by the eye-movement and the music video of our task. We treat this dataset as the ground truth and show the effectiveness of the proposed method by the measure for the evaluation which is proposed in [8]. The contributions in this paper are as follows. (1)We propose the saliency model for the task-driven eye-movement. (2)We add the information which is not the image feature as well as the image feature to the saliency model. (3)We introduce the musical information in the form of the musical note to the saliency model of the music video. (4)We treat the eye-movement which is observed in our tasks as the ground truth when evaluating the saliency model.

The rest of the paper is organized as follows. Section 2 is described the problem setting. Section 3 is described the proposed saliency model. Section 4 is described the dataset for the evaluation of the saliency model. Section 5 demonstrates experimental results. Section 6 is described the discussion and Section 7 is the conclusion.

2 Problems

Our goal is to detect the saliency for the estimation of the task-driven eye-movement. First, we describe the task setting, and then, we describe the saliency model based on the image feature that is baselines of our work.

2.1 Task Setting

Yarbus[1], Henderson[2], Angelusa[3] and Itti[4] gave the observers some tasks when observing a still image, and measured the eye-movement. In this paper, we gave the observers two tasks as follows in order to observe the task-driven eye-movements.

- Task 1: To evaluate the piano performance.
- Task 2: To memorize the music.

We consider that these tasks are affected by the musical discipline of the observers. So, the observers have the experience of learning to play the piano for more than one year in this paper. They are considered to have more chances of developing the musical sense than someone who has never learned the piano. The music videos in our work are related to the piano performance and collected from the video site "You Tube". We perform the previous measurement for our work. In the previous measurement, we measured the gaze behavior in watching the music video without observer's head fixed. The music video included whole body of the performer, and the video was taken from the right side of the performer. The music video also included the sound of the piano performance. The length of the video was 1 minute and the part of the tune was used. We found as follows in the previous measurement. First, since the observer's head was not fixed, the change of face direction and the range of movement of the observer's head was not restricted. We therefore consider that the accurate position of the gaze point was not obtained by the eye tracker. Second, most observers paid attention to the head of the performer. Third, observers tended to gaze at the center of the frame or a certain location as the time passed. According to the previous measurement, we selected the videos that meets the following conditions. (1)The music video includes the sound of the piano performance. (2)The music video was taken by the fixed camera. (3)The music video was taken from the position where the keyboard and the performer's hands are seen (Fig.1(a)). (4)The head of the performer is not seen in the video. (5)The person in the video is only the performer. There are no restrictions to the performer, music, the background of the video and piano. The performer basically play the piano according to the musical score. The observers also have difficulty in recognizing the motion of all of the fingers, so the ambiguity of the musical score and the sound is not considered in this paper. We use the eye-movement of the observers in two tasks as the ground truth to evaluate the saliency model. However, the ground truth is obtained by the eye-movement of some observers as is the case in [2][3][4].

2.2 The saliency map based on the image feature

There are many works where the saliency is detected by the image feature based on the saliency hypothesis 2. These works are divided into two types. One is the saliency model for the estimation of the fixation points from the natural images[6][9][10][11]. The other is the saliency model for the detection of the salient object in the image[7][12][13]. We use the saliency models proposed by Itti[6] and Yang[7] (the state-of-the-art) as the baselines and add the score feature to these baselines. We also describe each baseline. Itti constructed the three feature maps (intensity, color, orientation), and summed them to obtain the saliency map. Since we use the video, the optical flow is added as the dynamic feature. Yang constructs the graph where each superpixel extracted from the image is used as the node. First, these nodes are compared with the nodes of four sides of the image (the top, bottom, left and right of the image) as labeled background queries, and compute the salient nodes based on their relevances (*i.e.*, ranking score) to those background queries, so that the labeled maps of each side are obtained. Then, these four labeled maps are integrated to generated a saliency map. Second, the labeled foreground nodes are taken as saliency queries, and the saliency of each nodes is computed based on the relevance to foreground queries for the final map.

3 Our Approach

In this section, we describe the proposed saliency of the music video that includes the musical score information.

3.1 The saliency adding the musical score feature

In this paper, we consider the image feature extracted from the music videos and the score feature S extracted from the musical score corresponding to the music video frame. We use these features to construct the proposed saliency model of the music videos. We do not intend to match the musical score to the striking keys in the musical video, so we do by hand. The automation of matching the musical score to the striking keys in the video is possible by the digital piano.

The musical score feature:
We extract the score feature from the musical score. The method for extraction is as follows. In this paper, we use the musical score as the musical feature. First, we related the notes on the score to the striking keys by using the virtual keyboard. The virtual keyboard has 52 white keys whose size are $75pixels \times 5pixels$ and 36 black keys whose size are $25pixels \times 5pixels$. The position of keying is defined as the center of each key. We transform the position of the notes on the virtual keyboard to the position of the keyboard in the music video. We also generate the normal distribution whose mean is the center of each position of keying. We consider the view angle of the visual field (0-5 degrees from the center of the visual field), so we define the variance as 20 pixels of this normal distribution[14]. In this way, we can obtain the feature map of the musical note.

Our model based on Itti's method:
We construct the image feature map (the color feature map C, the orientation feature map O, the intensity feature map I, the dynamic feature map M) from the image as is the case in [6]. Next, we summed these image feature maps and the musical score feature map with the weighted linear combination as following formula,

$$Sal_{Itti+S} = a_1C + a_2O + a_3I + a_4M + a_5S, \tag{1}$$

wherein $a_i(i = 1, \ldots, 5)$ is the weight of each feature map (as shown in Table 2). Fig.2(a) is an example of the saliency calculated by Itti ' s method, Fig. 2(b) is an example of the saliency map where the optical flow is added as the dynamic feature, Fig. 2(d)(e) are examples of the proposed saliency maps where the musical score feature is added.

Our model based on Yang's method:
We also calculate the saliency by the Yang ' s method. In Yang's method, the superpixels are extracted from the image as the node, and the salient region is detected using the graph-based ranking score of each node. We add the musical score feature to the ranking score defined as following formula.

$$\boldsymbol{f^*} = (\boldsymbol{D} - \alpha\boldsymbol{W})^{-1}\boldsymbol{y}, \tag{2}$$

where $\boldsymbol{f^*}$ is the ranking value of each node, $W = [w_{i,j}]_n \times n$ is an affinity matrix of the nodes, $D = diagd_{11}, \ldots, d_{nn}$ is the degree matrix, where $d_{ii} = \sum_j w_ij$, and y_i is whether the $node_i$ is the query. We add the musical score feature to the ranking value. First, we split the musical score map into the superpixels that are the same as that of the frame of the music video. The resolution of the musical score map is the same as that of the music video frame. We also compute the average of each superpixel of the musical score map, so that we obtain the musical value MV for each superpixel. We use the normalized value of MV and compute a new ranking score as follows,

$$\boldsymbol{fnew_i^*} = f_i^* \cdot MV_i. \tag{3}$$

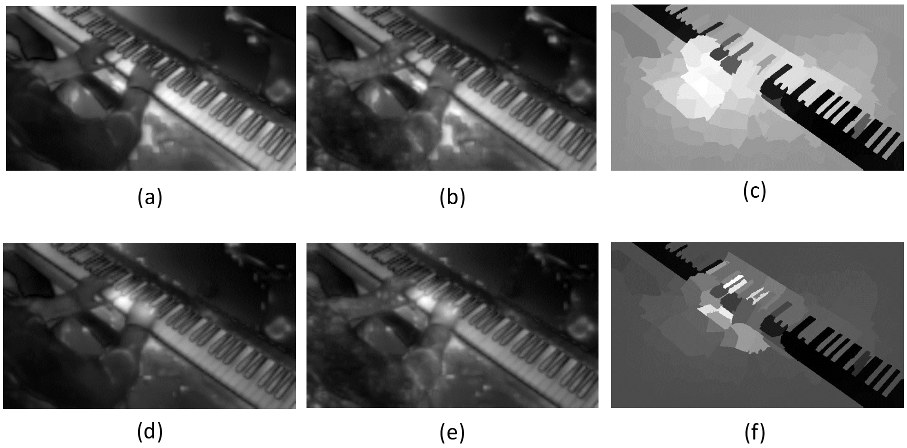

Fig. 2. The saliency map.(a) is the map proposed by Itti, (b) is the proposed map based on (a), (c) is the map which adds feature of the optical flow to (a), (d) is the proposed map based on (d) (e) is the map proposed by Yang, (f) is the proposed map based on (e)

We obtain the four labeled maps (the query of each map is the top, bottom, left and right of the image) and integrate them to obtain the saliency map based on the ranking score $fnew_i^*$. Fig. 2(c) is an example of the saliency calculated by baseline of Yang, Fig. 2(f) is an example of the proposed saliency map where the musical score feature is added.

4 DataSet

Our saliency model is task-driven, and the evaluation of our model needs the ground truth of the task driven eye-movements. For overt attention and the detection of eye fixation positions, the ground truth of human attentional behavior can be measured using an eye tracking device. The eye tracker provides the binocular gaze point at a joint sampling frequency of 60 fps. The framerate of the viewing video is 30 fps with a resolution of 640 × 480 pixels. The dataset has been proposed that includes the video and the eye tracking data. However, the task is not given to the observers in the existing dataset, and the video in the dataset does not include the sound. We therefore construct the dataset of the saliency in order to estimate the task-driven eye-movement. The observers consist of 12 observers who have learned the piano and 10 observers without the experience of a piano performance. The observer are 18 to 32 years old, both males and females. The number of years of learning the piano is 1 to 21. In addition, a professional artists is not contained in the observer. The time of

Table 1. The dataset of the gaze data.

Dataset contents

–The serial data (the frame counter, XY coordinate, data of gazing points, pupil diameter)
–The viewing image with gazing points
–The questionnaire

Subjects

–22 subjects of 12 experienced person and 10 inexperienced person

Video The Subjects Watched

–10 music videos(Each video is 30 seconds)
–Tune list(Number of tune, Name of classical composer, Title of tune, Degree of the visibility of the tunes)
No.1, Beethoven,Fur Elise， 2.425.
No.2,Chopin,Nocturne Op.9-2,1.93.
No.3,Beethoven,Piano Sonata No.17 ”Tempest”,1.27.
No.4,Rubinstein,Op.44,No1,1.04.
No.5,Schumann = Liszt, Widmung,1.14.
No.6,Chopin, Etudes Op.10-8,1.04.

each music video was about 30 seconds. We give the observers following two tasks. Task 1 is to evaluate the performance by five levels in four items (e.g, the mellifluence, the strength of sound, the accuracy of keying, the rhythm of the performance). Task 2 is to memorize the tune and select the score of the performed tune from three scores on the display after watching the music video. The observers watched eight videos in task 1 and two videos in task 2. Each video is 30 seconds. We measured the eye-movement of the observers in each task. The music videos were collected from ”You Tube” and met the conditions as described in Sec.2.1. The detail of the music videos is shown in Table 1. We also asked the visibility of each tune in three levels (ex.,unknown tune, known tune, played tune) to the observers. Since the musical sense developed in the musical discipline is considered in particular, we used the tune with high visibility among the observers. Additionally, we describe the generation of the ground truth of the task-driven eye-movements as shown in Fig. 3. The ground truth is the distribution of the gaze data of the observers in watching the music video, and generated per frame. We first overlap the visual fields of all the observers with the music video frame. Humans recognize the object in the viewing area centering on the gaze point. The area where human can visualize the object accurately is restricted because of the structure of the retina. The visual angle of this area is known as 0 to 5 degrees. However, the method for estimation of the

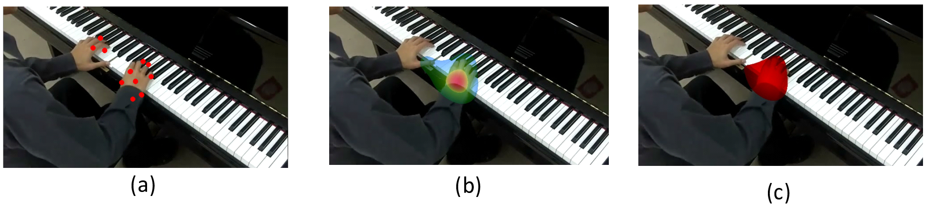

Fig. 3. The ground truth.(a)The gaze position is obtained per frame. (b) The visual fields overlapped on the music video frame. (c) The ground truth obtained by binarization of the map(b).

accurate visual fields individually has not been established yet, so many works define the visual fields as the circular area. We also approximate the visual fields as follows,

$$R_{cm} = d \times tan(\frac{\theta}{2} \times \frac{\pi}{180}), \tag{4}$$

$$R_{px} = R_{cm} \times \frac{w_{px}}{w_{cm}}, \tag{5}$$

where R_{cm} and R_{px} represent the radius of the visual fields in centimeters and in pixels, θ represents the visual angle of the visual fields, w_{cm} represents the width of the display in centimeters, w_{px} represents the width of the music video frame in pixel. In our work, θ was $5degree$, w_{cm} was 59.79 cm, w_{px} was 1920 pixels. From these parameters, R_{px} was 59 pixels. The visual fields of all the subjects were overlapped with one frame and binarized, so that we can obtained the ground truth. The threshold for binarization was 0.7 of the maximum of the overlapped frame[8].

5 Experiments

We evaluate three baselines and three our methods by comparison between the saliency map and the ground truth per frame of the music video. We use the tune with high visibility that is well-known among the observers.

5.1 Evaluation metric

For the evaluation of the saliency map, we used Normalized Scanpath Saliency (NSS)[15], the Correlation Coefficients (CC)[16], the area under the Receiver Operating Characteristics (AUC-ROC)[17], and precision, recall, F-measure as the evaluation value.

Normalized Scanpath Saliency(NSS)

For computing the NSS value, the saliency map was linearly normalized to have zero mean and unit standard deviation. NSS is obtained as following,

$$NSS = \sum_{i=1}^{n} \frac{s(x_i^h, y_i^h) - \mu_s}{\sigma_s}, \tag{6}$$

where $s(x_i^h, y_i^h)$ is the normalized saliency value of the locations of the ground truth, μ_s and σ_s are the mean and variance of the normalized saliency map. A value greater than zero suggest that the saliency map correspond to the eye position of the ground truth, a value of zero indicates no correspondence between the saliency map and the ground truth, and a value less than zero indicate an anti-correspondence.

Correlation Coefficients(CC)
CC is obtained as follows,

$$CC(s, h) = \frac{cov(s, h)}{\sigma_s \sigma_h}, \tag{7}$$

where s is the map of the ground truth, h is the saliency map, σ_s and σ_h are the variance of each map. The value close to 1 indicates that the saliency map correspond to the ground truth map, the value close to 0 indicates no correspondence between the saliency map and the ground truth, and the value close to -1 indicates an anti-correspondence.

the area under the Receiver Operating Characteristics(AUC-ROC)
ROC curve is the signal detection theory. First, fixation pixels of the ground truth are positive set, and the same number of random pixels are chosen from the saliency map as the negative set. The saliency map is then treated as a binary classifier, and all points above threshold indicates positive samples and all points below threshold indicates negative samples. By plotting true positive rate vs. false positive rate for any particular value of the threshold, an ROC curve can be drawn and the Area Under the Curve(AUC) computed. An ideal score is one while random classification provides 0.5.

Precision, Recall, F-measure
We use precision, recall and F-measure for the evaluation. In order to calculate these values, we determined the adaptive threshold to binarize the saliency map. The threshold is obtained as follows[18],

$$T_\alpha = \frac{\alpha}{W \times H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} S(x, y). \tag{8}$$

The adaptive threshold is α times the mean saliency of the music video frame. α is 1 to 5 and we adopted the F-measure value where the mean of F-measure of 6 saliency models is the highest in the level of α. F-measure is described as follows.

$$F_\beta = \frac{(1 + \beta^2) Precision \times Recall}{\beta^2 \times Precision + Recall}. \tag{9}$$

We use $\beta^2 = 0.3$ to weigh precision more than recall[18].

5.2 Result

Table 2. The weight of each saliency model based on Itti's method

	Parameter				
	Intensity	Color	Orientation	Motion	Score
Itti(Baseline)	0.33	0.34	0.33	0	0
Itti+M(Baseline)	0.3	0.3	0.2	0.2	0
Itti+S(w1)	0.3	0.3	0.2	0	0.2
Itti+M+S(w1)	0.25	0.25	0.15	0.2	0.15
Itti+S(w2)	0.25	0.25	0.25	0	0.25
Itti+M+S(w2)	0.2	0.2	0.2	0.2	0.2
Itti+S(w3)	0.2	0.2	0.2	0	0.4
Itti+M+S(w3)	0.15	0.15	0.15	0.15	0.4

In this section, we evaluate the baseline models and our proposed models by the evaluation values above stated. As the ground truth, we used the gaze behavior when watching the part of the tune ”Fur Elise (Beethoven)” as shown in Table 1 whose visibility was high among the observers. We also set the weight of our saliency models based on Itti's method as shown in Table 2. From the result shown in Table 3, our models adding the musical score feature outperform the baselines that are constructed from the image feature. From the 7th to 10th row in Table 3 are our saliency models that weights the music score map more than the image feature as shown in Table 2. From the result, the evaluation values increase as the weight of the musical score map is large. Additionally, our saliency model of ”Itti+M+S(w3)” outperform the score map in most values. This indicates that we need not only the musical score feature but also the image feature in order to construct the saliency map of the music video. As for the threshold (the formula 8), the F-measure value is the highest when α is 3. We therefore use that threshold for the following evaluation.

6 Discussion

In this section, we also evaluate the saliency map under some condition, and discuss the result. First, we generate the ground truth from the gaze behavior of the inexperienced subjects. In our work, we consider that everyone has the possibility of having the musical sense. We therefore measured the gaze behavior of the inexperienced subjects in watching the music video, and evaluated the saliency maps using that ground truth. The result is shown in Table 4. Compared to the baselines that are constructed from the image feature, our saliency models that are added the musical score feature give high values. Our saliency model based on Itti ’ s method gives high values when the musical score feature map is weighted more. From this result and Table 3, we consider that the musical score

Table 3. Result of the evaluation of the saliency maps

	AUCROC	CC	NSS	F-measure	Precision	Recall
Itti(Baseline)	70.3%	0.105	0.65	6.91%	6.57%	9.64%
Itti+M(Baseline)	73.6%	0.127	0.779	1.8%	2.13%	1.38%
Yang(Baseline)	84.2%	0.299	1.85	9.8%	10.8%	11.5%
Itti+S(w1)	84.2%	0.24	1.49	19.4	21.1%	17.8%
Itti+M+S(w1)	86.1%	0.256	1.59	18.5%	27%	10.7%
Yang+S	76.7%	0.174	1.03	5.35%	9.21%	2.8%
Itti+S(w2)	84.7%	0.259	1.61	22.2%	21.8%	26.3%
Itti+M+S(w2)	86.8%	0.280	1.74	24.1%	**27.1%**	20.2%
Itti+S(w3)	85.8%	0.321	2.23	24.3%	24.2%	40.1%
Itti+M+S(w3)	**87.7%**	**0.354**	**2.22**	**26.5%**	24.3%	**41.1%**
Score Map	75.2%	0.355	2.23	22.0%	18.6%	60.0%

Table 4. The result of the evaluation of the saliency maps (the ground truth is generated from the gaze behavior of the inexperienced subjects).

	AUCROC	CC	NSS	F-measure	Precision	Recall
Itti(Baseline)	67.7%	0.078	0.47	3.3%	3.3%	4.4%
Itti+M(Baseline)	73.1%	0.11	0.70	0.9%	1.3%	0.6%
Yang(Baseline)	85.2%	0.28	1.82	5.5%	5.3%	9.1%
Itti+S(w1)	82.5%	0.21	1.34	16.8%	18.4%	15.2%
Itti+M+S(w1)	86.1%	0.24	1.56	17.6%	**25.3%**	10.4%
Yang+S	77.3%	0.16	1.00	3.3%	6.3%	1.9%
Itti+S(w2)	82.7%	0.23	1.47	19.1%	18.9%	22.8%
Itti+M+S(w2)	86.5%	0.26	1.69	22.0%	24.5%	19.4%
Itti+S(w3)	83.8%	0.29	1.87	22.9%	21.2%	36.8%
Itti+M+S(w3)	**87.1%**	**0.34**	**2.16**	**24.1%**	22.2%	**39.7%**
Score Map	74.5%	0.34	2.18	20%	17.0%	58.6%

Table 5. The result of the evaluation of the saliency map. Case 1 is the result that the saliency map includes the musical information of the current tunes and a subsequent tunes. Case 2 is the result that the saliency map includes the musical information of the current tunes, a subsequent tunes and a previous tunes. Case 3 is the result that the saliency map includes the current tunes and 2 subsequent tunes. Case 4 is the result that the saliency map includes the current tunes, 2 subsequent tunes and 2 previous tunes.

Case	Method	AUCROC	CC	NSS	F-measure	Precision	Recall
1	Itti+S(w3)	88.1%	0.33	2.02	24.0%	21.9%	38.5%
	Itti+M+S(w3)	89.6%	0.36	2.23	**24.5%**	**22.2%**	40.9%
	Yang+S	**90.1%**	0.27	1.63	11.0%	9.2%	39.2%
	Score Map	78.0%	**0.38**	**2.32**	21.5%	19.1%	**66.8%**
2	Itti+S(w3)	88.5%	0.33	2.05	24.5%	22.6%	38.1%
	Itti+M+S(w3)	**90.2%**	0.37	2.28	**25.3%**	**23.3%**	39.4%
	Yang+S	88.4%	0.26	1.55	10.9%	9.1%	39.6%
	Score Map	82.5%	**0.38**	**2.39**	23.0%	19.1%	**77.6%**
3	Itti+S(w3)	84.8%	0.28	1.68	19.6%	18.2%	29.0%
	Itti+M+S(w3)	86.7%	**0.31**	**1.87**	**20.7%**	**19.2%**	31.3%
	Yang+S	**88.6%**	0.27	1.62	13.4%	11.0%	54.2%
	Score Map	72.4%	0.29	1.75	18.6%	15.6%	**56.4%**
4	Itti+S(w3)	85.2%	0.28	1.68	18.4%	17.0%	28.5%
	Itti+M+S(w3)	87.5%	**0.31**	**1.86**	**19.2%**	**17.6%**	29.9%
	Yang+S	**89.5%**	0.25	1.54	9.9%	8.1%	40.9%
	Score Map	73.9%	0.29	1.76	18.0%	15.0%	**59.0%**

feature other than the image feature can detect the saliency that is close to the observed gaze behavior when watching the music video. The evaluation value of the Table 4 tends to be lower than that of the gaze behavior of the experienced subjects. We consider that the experienced subjects watched the piano performance with attention to the performer's hands and the tune. Regardless of the experience of learning the piano, the recall of the musical score map and the saliency map where the musical score map are weighted is high, which shows that the area where the ground truth correspond to the score feature map is large, but the salient area detected by the image feature does not correspond to the ground truth.

Second, we generated the musical score map for four patterns, where the number of notes are different as follows,

- Case 1: The score feature is extracted from the striking keys and one note after the striking keys.
- Case 2: The score feature is extracted from the striking keys and one note before and after the striking keys.
- Case 3: The score feature is extracted from the striking keys and one and two notes after the striking keys.
- Case 4: The score feature is extracted from the striking keys and one and two notes before and after the striking keys.

In each case, we weigh the striking keys more than other notes. We use these musical features to construct our proposed saliency models. The result of the evaluation is shown in Table 5, where the ground truth is generated by the gaze behavior of the experienced subjects. The evaluation values of Case 1 and Case 2 outperform the result of Table 4. We therefore find that we can obtain more appropriate model to the eye-movements in our tasks by adding the musical notes other than the striking keys as the musical score feature. However, most of the evaluation values in Case 3 and Case 4 are lower than that of Case 1 and Case 2. To add the musical notes as the musical score feature leads to expand the salient area, which increases the area corresponding not to the ground truth.

7 Conclusion

We proposed the task-driven saliency model of the music video. We added the musical score map generated from other than the image feature, and constructed the saliency model based on the baselines constructed from the image feature. The proposed saliency models were evaluated by comparison to the ground truth that was generated the gaze behavior when watching the music video, and we showed that the musical score feature as well as the image feature are needed for detecting the saliency of the music video. In this paper, we used the musical notes as the musical information. However, other information related to the musical sense is possibly preferable to detect the task-driven saliency of our tasks, such as the sound intensity, the rhythm, and so on. Additionally, the ground truth was generated by the observer having learned the piano more than one year because we consider that they have relatively many chances of developing the musical sense. However, we should consider other information that is related to the musical information except the experience of learning the piano.

References

1. A. L. Yarbus, Eye movements and vision, New York,Plenum Press,1967.
2. J.M. Henderson, S.V. Shinkareva, J.Wang, S.G.Luke and J.Olejarczyk, Predicting cognitive state from eye movements, Plos One, 8(5), e64937,2013.
3. M.DeAngelusa and J.B.Pelza, Top-down control of eye movements: Yarbus revisited, Visual Cognition,Vol.17, pp.790-811, 2009.
4. A. Borji and L. Itti, Defending Yarbus: Eye movements reveal observers' task, Journal of vision,2014.
5. K.Kunze, Y.Utsumi, Y.Shiga, K.Kise and A.Bulling, I know what you are reading: recognition of document types using mobile eye tracking, International symposium on wearable computers, 2013.
6. L.Itti, C.Koch and E.Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, PAMI,Vol.20, No.11, pp.1254-1259,1998.
7. C.Yang, L.Zhang, H.Lu, X.Ruan and MH.Yang, Saliency Detection via Graph-Based Manifold Ranking, CVPR, pp.3166-3173, 2013.
8. N.Riche, M.Mancas, D.Culibrk, V.S.Crnojevic, B.Gosselin and T.Dutoit, Dynamic saliency models and human attention: a comparative study on videos, ACCV, Vol.7726, pp.586-598, 2012.
9. N. Bruce and J. Tsotsos, Saliency based on information maximization. NIPS, 2005.
10. J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. NIPS, 2006.
11. H.J.Seo and P.Milanfar, Static and Space-time Visual Saliency Detection by Self-Resemblance, The Journal of Vision,Vol.9,No.15,pp.1-27, 2009
12. L. Wang, J. Xue, N. Zheng, and G. Hua. Automatic salient object extraction with contextual cue, ICCV, 2011
13. K.Shi, K.Wang, J.Lu and L.Lin, PISA: Pixelwise Image Saliency by Aggregating Complementary Appearance Contrast Measures with Spatial Priors, CVPR, 2013.
14. A.Iwatsuki, T.Hirayama and K.Mase "Analysis of Soccer Coach's Eye Gaze Behavior", Proc. ASVAI, 2013.
15. Peters, Robert J., et al. "Components of bottom-up gaze allocation in natural images." Vision research 45.18 (2005): 2397-2416.
16. N. Ouerhani, R. von Wartburg, H. Hugli, and R.M. Muri. Empirical validation of saliency-based model of visual attention. Electronic Letters on Computer Vision and Image Analysis, 2003. 5
17. Bruce, Neil, and John Tsotsos. "Saliency based on information maximization." Advances in neural information processing systems. 2005.
18. R.Achanta, S.Hemami, F.Estrada, and S.Susstrunk, " Frequency-tuned salient region detection, " CVPR, 2009.