An Intelligent Captioning System for the Optimization of Image and Caption Visibility

Yuichiro Kinoshita, Ryoichi Shimizu, Eric W. Cooper, Yukinobu Hoshino and Katsuari Kamei

Department of Human and Computer Intelligence, Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577, JAPAN
Email: {kino, shimizu, eric, hoshino, kamei}@spice.ci.ritsumei.ac.jp

Abstract – **Several TV programs offer closed captions for hearing impaired people these days. However current closed captions overlap background image and sometimes have problems in terms of its visibility. In this paper, we propose an intelligent captioning system that adjusts the lightness of captioning areas in real time to optimize image and caption visibility. First we conducted visibility evaluation experiments to find out appropriate lightness values for the captioning areas. Based on the results, the system was constructed using a neural network. After the construction, we performed system evaluation. The results show that our system enhanced caption visibility without reducing image visibility.**

Keywords – **closed captions, intelligent systems, neural network, optimization, visibility**

I Introduction

Some TV programs offer closed captions these days. Not only for foreign language broadcasts, closed captions are effective way to provide information for hearing impaired people or foreign language speakers.

The normal closed caption is usually shown with white characters on a black background. The image behind the caption is not visible in this case. For this problem, the reduction in lightness of a captioning area is commonly used. However, this lightness offset is usually fixed at a certain value and it is difficult to determine one appropriate value for an entire program. The captioning area sometimes becomes too dark to recognize the background image.

On the other hand, there is another method in which an image is scaled down and a frame for captions is made on the edge of a screen [1]. Captions and an image don't overlap in this method, but the image becomes smaller and its visibility is reduced. Furthermore it is impossible to allocate captions on various positions. Methods of captioning that preserve both image and caption visibility are expected to improve overall viewing quality.

In this paper, we propose an intelligent captioning system that adjusts the lightness of the captioning area appropriately in real time to optimize both image and caption visibility. The system is constructed with a neural network trained by visibility evaluation experimental results on human subjects.

II Visibility Evaluation Experiments

A Lightness adjustment method

In this research, we express colours using the HLS colour space [2], in which every colour is represented by three attributes: hue, lightness and saturation. Lightness of colours will be adjusted by the multiplication of the scale α. We define the pixel located at (x, y) in a captioning area on the image d as $P_d(x, y)$ when the pixel at the upper left corner is (0, 0). Here the lightness value after the adjustment can be expressed as

$$L_d^{'}(x, y) = \alpha \times L_d(x, y) \qquad (0 \leq \alpha \leq 1) \tag{1}$$

where $L_d(x, y)$ represents lightness value of $P_d(x, y)$ in the HLS colour space.

When only lightness value is adjusted, the look of saturation also varies due to features of the HLS colour space. The saturation value also needs to be adjusted concurrently with lightness value in order to keep the saturation looking the same. The saturation after the adjustment can be defined as

$$S_d^{'}(x, y) = S_d(x, y) \times \frac{1.0 - L_d(x, y)}{1.0 L_d^{'}(x, y)} \tag{2}$$

where $S_d(x, y)$ represents saturation value of $P_d(x, y)$ in the HLS colour space.

B Evaluation method

Visibility evaluation experiments are conducted to find appropriate lightness values for the captioning area. Ten subjects including two female participated in the experiments. A 30-inch flat screen television was used for the experiments and the subjects sit 2.5m away from the television. Fig. 1 shows the screen layout for the experiments. A closed caption is placed on the lower part of the screen with white round-gothic font. The caption is shown with two lines and each line consists of fourteen Japanese characters.

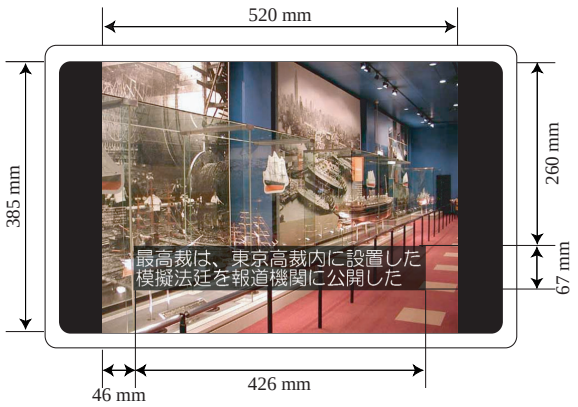

Fig. 1 The experiment screen layout.

150 picture images, such as scenery, buildings or people, were chosen as the samples. For the background images, the subjects adjust the lightness of the captioning area and decide an appropriate value for the scale α to have good visibility for both image and caption. The subjects were allowed to modify the lightness adjustment scale α in steps of 0.033 using a keyboard. Here α determined by each subject may be depending on the initial value of α. Hence we employed the following two methods in this time.

Method 1 The subjects start the experiments with α=1 and decrease α step by step.

Method 2 The subjects start the experiments with α=0 and increase α step by step.

The subjects evaluated all of 150 images. The images were shown in random sequence to reduce influence from the order of the presentation. The subjects took a thorough break every 25 samples.

C *Determination of lightness adjustment scale*

An appropriate lightness adjustment scale is determined for each image based on the experiment results.

Here we define the lightness adjustment scale α_d for the image d decided by the subject $s(s = 1, 2, \cdots, N)$ using the method 1 or the method 2 as α^1_{sd} or α^2_{sd}, respectively. We compared α^1_{sd} with α^2_{sd} and there was large difference among them for some cases. This kind of results don't have enough reliability since subjects are supposed to determine α_d based on the same basis. At this time, we excluded results in which $|\alpha^1_{sd} - \alpha^2_{sd}| > 0.33$.

We determined the lightness adjustment scale α_d for the image d as

$$\alpha_d = \frac{1}{N}\sum_{s=1}^{N} \max(\alpha^1_{sd}, \alpha^2_{sd}) \tag{3}$$

where $N = 10$ in case of the experiments.

Fig. 2 shows the determined lightness adjustment scale α_d for all sample images. Those results show that

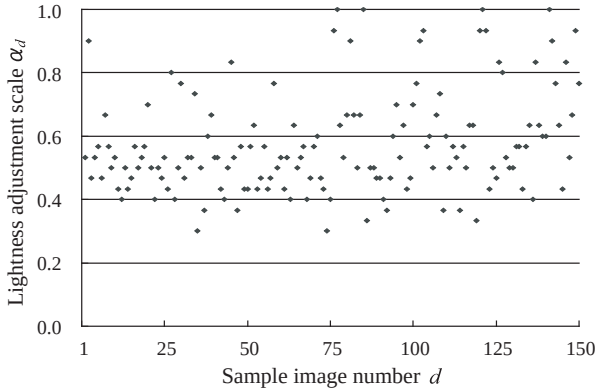

Fig. 2 Lightness adjustment scale α_d determined by subjects.

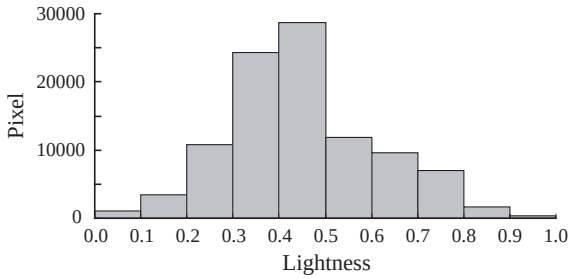

Fig. 3 An example of lightness histogram.

an appropriate α value providing good visibility differs depending on the lightness value of the captioning area.

III Intelligent Captioning System

A *Extraction of feature vector*

A lightness histogram of the captioning area colours was made for every sample image. For the histogram, lightness axis [0, 1] was divided into ten intervals and all pixels in the captioning area were classified depending on their lightness values. Fig. 3 shows an example of the lightness histograms.

Based on the lightness histogram, we considered the vector

$$\boldsymbol{H}_d = \{\ell_{da} | a = 1, 2, \cdots, 10\} \tag{4}$$

where ℓ_{da} represents the proportion of the number of pixels counted for the interval a to the total number of pixels in the captioning area.

Factor analysis with the principal factor method was performed using $\boldsymbol{H}_d$ for the 150 image samples. A factor were extracted where the eigen value was over 1.0. In this case, four factors were extracted in total. Table 1 shows the result of the factor matrix rotated by Varimax method [3]. Here we chose seven variables from the variables of $\boldsymbol{H}_d$ where one of factor loadings was over 0.7. Finally the feature vector of the captioning area colours was extracted as

$$\boldsymbol{F}_d = \{\ell_{d2}, \ell_{d3}, \ell_{d5}, \ell_{d6}, \ell_{d7}, \ell_{d8}, \ell_{d9}\}. \tag{5}$$

Table 1 Factor loading matrix.

	Factor 1	Factor 2	Factor 3	Factor 4
ℓ_{d8}	**0.81**	0.12	-0.02	-0.09
ℓ_{d9}	**0.81**	-0.04	0.06	-0.10
ℓ_{d6}	-0.12	**-0.79**	0.35	-0.15
ℓ_{d7}	0.26	**-0.79**	0.03	-0.06
ℓ_{d2}	-0.33	-0.15	**-0.79**	0.00
ℓ_{d5}	-0.22	0.23	**0.78**	0.02
ℓ_{d3}	-0.21	-0.19	-0.13	**0.82**
Eigen value	2.51	2.10	1.38	1.03
Accumulated contribution	25.14	46.17	60.05	70.38

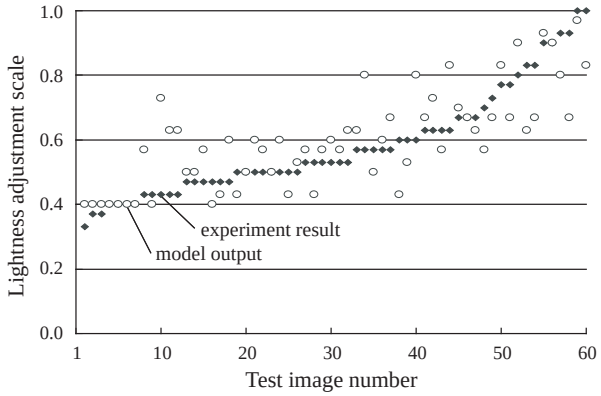

Fig. 4 The testing results.

B *Lightness adjustment model*

Based on the experiment results, we modelled the relationship between appropriate lightness and colours of the captioning area using neural network. We chose 90 experiment results as a training data set for the neural network and remaining 60 results were used as a testing data set. The data in each set were selected to have various α_d values.

The feature vector $\boldsymbol{F}_d$, which is extracted based on the lightness histogram, is input to the neural network and the output is a determined appropriate lightness value for the input sample. We used the back-propagation as a learning algorithm.

After the construction, we tested the performance of the neural network. We input the 60 testing samples into the constructed model and calculated errors between the output values from the model and the experiment results. The testing results showed that the error was less than 10% for over 75% of the samples as shown in Fig. 4. The results give a proof that this network have adequate ability to determine appropriate lightness for captioning areas.

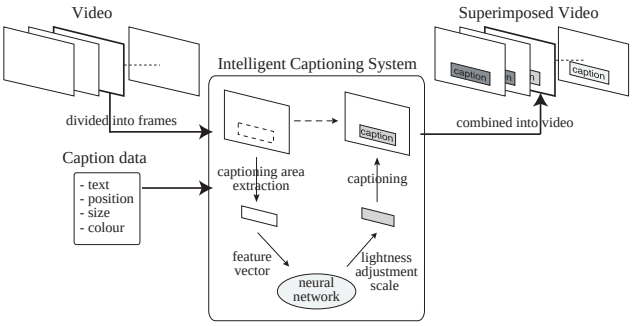

Fig. 5 The components of intelligent captioning system.

C *Implementation of intelligent captioning system*

The intelligent captioning system was constructed using the trained neural network. Input items to the system are video data as well as caption data including captioning position and font size. Output item is superimposed video data. Fig. 5 shows the components of the system.

In the system, the input video data is divided into 30 frames per second and a captioning area will be determined based on the input caption data. For each frame, a feature vector $\boldsymbol{F}_d$ is generated and then the neural network finds appropriate lightness adjustment scale α_d. The system adjusts the captioning area lightness and superimposes the caption. Finally the system combines the frames and outputs superimposed video data.

IV User Evaluation of Intelligent Captioning System

A *Methods*

After the system construction, we compared the performance of our system with the following two types of closed captions.

Caption A White characters with fixed background lightness (fixed at 0.7)

Caption B White characters with 2 pixel black outline (no background lightness control)

We performed two types of evaluation for eight participants. Ten portions of video from news programs are used as evaluation samples and several closed captions are sequentially superimposed on the video. Each video is around 45 seconds. The closed captions are changed in 5.4 second cycle. One closed caption is shown for 4.7 seconds and then cleared. 0.7 second later, another caption is shown up. A 30-inch flat screen television was used for the evaluation and the participants sit 2.5m away from the television. The screen configuration is the same as Fig. 1.

The first evaluation is based on questionnaires. The participants evaluate the captions and the background video using five-step Likert scale in terms of visibility.

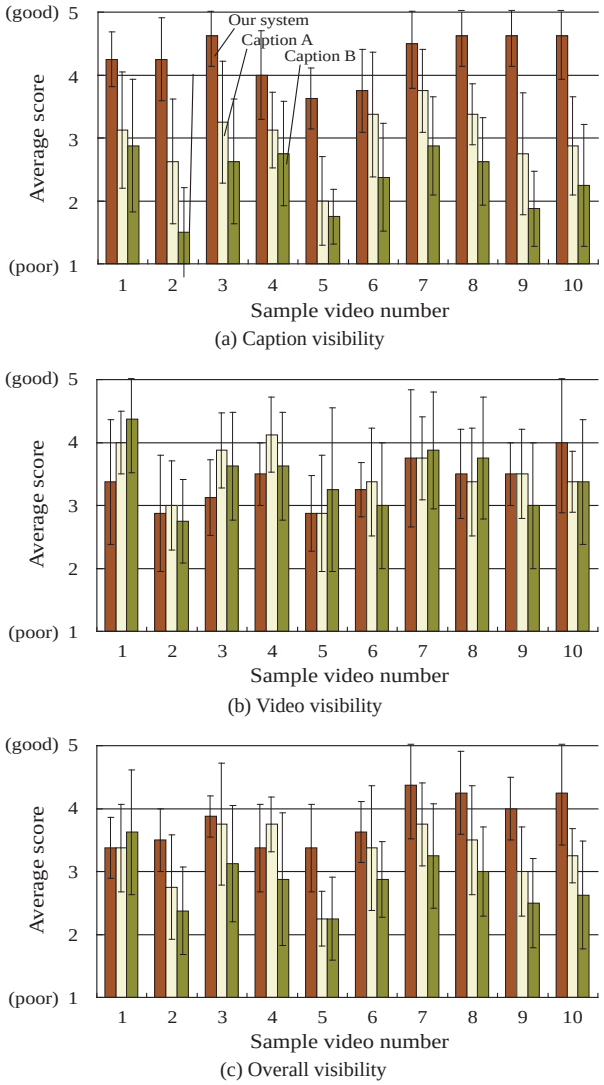

Fig. 6 The results of the questionnaire evaluation.

For the second evaluation, the participants watch the video and keep pressing a key on a keyboard while they feel that the caption has poor visibility.

B Results

The results of the questionnaire evaluation are shown in Fig. 6 where 1 represents poor visibility and 5 represents good visibility. The vertical lines in the figures represent standard deviation. For the caption visibility evaluation, there are significant differences between our system and the others as shown in Fig. 6(a). On the other hand, there is no significant difference between the three for video visibility as shown in Fig. 6(b). These results show that our system achieved the enhancement of caption visibility without reducing video visibility. We also performed the overall visibility evaluation as shown in Fig. 6(c). The score showed that the visibility of our system was evaluated as the best for 80% of the samples and the second-best for the rest of the samples.

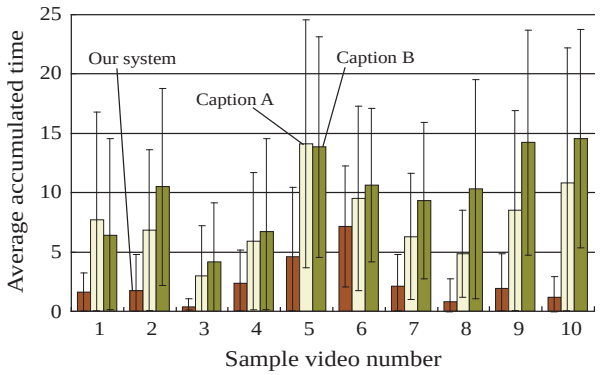

Fig. 7 The accumulated time period evaluated as poor visibility.

Fig. 7 shows the accumulated time period while the participants felt that the caption has poor visibility. The result showed that the total time evaluated as poor visibility in our system was 69% shorter than Caption A and 76% shorter than Caption B. For most video samples, the results of standard deviation showed large value. It means the participants made their decisions based on their own basis. However, among all participants, the time period evaluated as poor visibility in our system was obviously shorter than those in Caption A and B.

V Conclusions

In this paper, we constructed an intelligent captioning system. The system adjusts the lightness value of captioning areas appropriately in real time. First we conducted visibility evaluation experiments to find out appropriate lightness values for the captioning area. Based on the experiment results, we constructed the system using the neural network. After the construction, system evaluation was performed. The results show that our system enhanced caption visibility and reduced the time while they feel that the caption has poor visibility.

As our future work, we will expand our system to handle video having specific patterns and/or colours such as animation programs.

References

[1] Monma,T., Sawamura,E., Mitsuhashi,T., Ehara,T., Shirai,K.: Subjective Evaluation of Methods for Presenting Closed-Captions on TV News Programs for Hearing-Impaired People. The Journal of the Institute of Image Information and Television Engineers, Vol. 54, No. 9, pp. 1288–1297 (2000) (in Japanese).

[2] Russ, J. C.: The image processing handbook (second edition). CRC Press (1995).

[3] Kaiser, H. F.: Varimax solution for primary mental abilities. Psychometrika, Vol. 25, pp. 153–158(1960).