IIEEJ **Paper**

Face Detection for Comic Images Using the Deformable Part Model

Hideaki YANAGISAWA†, Daisuke ISHII†† *(Member)*, Hiroshi WATANABE† *(Member)*

†Graduate School of Fundamental Science and Engineering, Waseda University, ††Fujitsu Laboratories Ltd.

<Summary> In this study, the performance of character face detection using the Deformable Part Model (DPM) for comic images is investigated. Character faces in comics are an important metadata to realize highly functional e-comics. The performance of DPM is compared with Histograms of Oriented Gradients (HOG), and greater detection performance is obtained in terms of precision-recall characteristics. Furthermore, side views and occluded samples of character faces are treated as detection targets. We apply DPM using multiple root filters and examine how the detection rate changed with the number of root and part filters. From the experimental result, the combination of 10 root filters and 11 part filters provides the highest average precision for character faces.

Keywords: deformable part model, HOG, comic, face detection

1. Introduction

Publishing has shifted from paper-based to electronic versions because of the rapid development of digital handheld devices such as tablets and smartphones. Within electronic content, the sales volume of e-comics accounted for more than 80 % of electronic publications in 2011 in Japan[1].

At present, most e-comic files consist of images, which may be in JPEG or PDF formats, with little or no tags. Therefore, it is difficult to search and access a designated page within a comic. A highly functional, searchable e-comic with tags is desirable. However, it is difficult to define metadata manually. Therefore, automatic metadata extraction from comics is necessary.

Information related to comic characters is an important metadata. A comic character may be a virtual human, an animal, or a creature. Once characters in a comic are detected and identified, character tags can be used to create a comic digest to search for similar comics at the reader's convenience. It is well known that Haar-like features are a powerful tool to detect humans. However, Haar-like features are not suitable for comics because images in comics are binary and parts of character faces are deformed and sometimes omitted. Instead, HOG features are thought to be suitable tools for comics. However, HOG's detection performance has not yet been sufficient for use in comics[2)-3)].

Researchers have shown that the Deformable Part Model (DPM) provides much higher character detection performance for comics than HOG[4]. However, the following issues have remained unsolved.

(1) The number of training images is small.

(2) Only front faces of characters are used.

(3) Faces with occlusions are omitted in the testing.

(4) Only the average values of recall, precision, and F-measure are provided.

In this study, we experiment with character face detection using DPM and focus on the following points. We prepared a sufficient number of training and testing images, including front faces, side faces, and front faces with partial occlusions. The state-of-the-art technology proposed by Orozco et al. in DPM can treat side faces[5]. We tried to optimize the number of root filters because Orozco's approach allows the usage of multiple root filters, and we further optimized the number of part filters because the original DPM was designed to detect a human body.

The rest of this paper is structured as follows. In Section 2, we briefly survey previous studies. In Section 3, we describe the features of comic images. In Sections 4, we describe multi-view face detection method, which are essential technologies in this paper. The proposed method for character face detection is described in Section 5. In Sections 6 and 7, we present and discuss the experimental results. Finally, we conclude in Section 8.

2. Previous Studies

Several metadata extractions from comics have been studied for frames (panel layouts), balloons, and characters. Ishii and Tanaka proposed a method to identify frames by obtaining frame boundary lines from the directions of the gradient's intensity along the boundary lines[6)-8)]. Nonaka proposed a method to detect rectangular regions and identify frames. In both techniques, the detection rate exceeded 80 %[9]. For balloon detection, Tanaka proposed a method that identifies word regions using AdaBoost, lists possible balloon candidates, and classifies the shape of the balloons using Support Vector Machine (SVM)[10]. Using this approach, it was reported that 86 % of balloons could be identified. For character extraction, Ishii and Arai proposed methods to detect face candidate regions via HOG features and identify characters by comparing candidates and faces in a database [2] [3]. HOG features are feature descriptors proposed by Dalal et al. for human detection[12]. HOG features consist of feature vector histograms of the luminance gradient intensity in a local area. They are not greatly affected by illumination changes and are robust to local geometry changes. HOG features are calculated using the following steps.

(1) Calculate the luminance gradient direction and the strength of the luminance intensity for each pixel in the image. Next, discretize the gradient direction into nine directions.

(2) Split a local area into cells, where each cell consists of 8 × 8 pixels.

(3) Create a luminance gradient direction histogram for each cell area and obtain nine-dimensional gradient vectors.

(4) Combine the gradient vectors in one block (2 × 2 cells) and normalize the luminance gradient histogram for the block. In this way, we obtain a multi-dimensional feature vector composed of the gradient vectors of all the blocks.

However, the facial parts of comic characters can dramatically vary depending on the scene. Deformation is often used in comics, e.g., exaggerated facial expressions. Therefore, stable character detection in comics using HOG features is not efficient.

DPM is an object detection method proposed by Felzenszwalb et al.[13] [14] DPM expresses an object as a model with several parts. The model is evaluated by the shape of the entire object and each part and the position movement of the parts. Therefore, DPM has structural characteristics using image features, which allows for deformed parts. From this viewpoint, DPM is regarded to be suitable for comic character detection.

The detection model in DPM consists of a root filter and part filters. The root filter captures the rough shape of the entire object. The part filters capture the shapes of each movable part of an object. DPM uses HOG features for its feature descriptors. HOG features are calculated per 8 × 8 pixels in multiple resolution images and are combined as a HOG pyramid. The cell position is expressed as $p = (x, y, l)$, where l indicates the resolution level of the HOG pyramid. At any position p in the HOG pyramid H, the vector-connected HOG features in the $w \times h$ blocks are expressed as $\phi(H, p, w, h)$. The root filter coefficient is expressed as F and the filter output as $F \cdot \phi(H, p, w, h) = F \cdot \phi(H, p)$. Here the root filter coefficient is represented as F_0, and the *n*-part filter is represented as $F_1, \cdots F_n$. The evaluation function of a score at a certain point $z = (p_0, p_1, \cdots, p_n)$ is given by Eq.(1).

$$
\begin{aligned}
score(p_0, \cdots, p_n) = & \sum_{i=0}^{n} F_i \cdot \phi(H, p_i) \\
& - \sum_{i=1}^{n} d_i \cdot \phi_d(dx_i, dy_i) + b
\end{aligned}
\tag{1}
$$

The displacement of the *i*-th part relative to its anchor position is given by Eq.(2),

$$(dx_i, dy_i) = (x_i, y_i) - (2(x_0, y_0) + v_i) \tag{2}$$

while the deformation features are described in Eq.(3),

$$\phi_d(dx_i, dy_i) = \left(dx_i, dy_i, dx_i^2, dy_i^2\right) \tag{3}$$

Detection model matching is performed by the following process: (1) Calculate the HOG pyramid from multiple resolution images, (2) Calculate the responses of the root filter and part filters, (3) Calculate the response of the part filters considering the cost of the movement, (4) Take the sum of all the filters' scores from Eq.(1), and (5) Detect the location where the score is larger than the threshold.

DPM uses Latent Support Vector Machines (Latent SVM) to estimate the model parameters for learning. The discriminant function for variable *x* is defined in Eq.(4):

$$f_\beta(x) = \max_{z \epsilon Z(x)} \beta \cdot \Phi(x, z), \tag{4}$$

where β is a model parameter vector and z are the latent values. In DPM, latent variables indicate part positions $(p_1, \cdots, p_n)$. As with normal SVM, we can calculate the parameters β for a given dataset $D = ((x_1, y_1), \cdots, (x_N, y_N))$, where $y_i \epsilon \{-1,1\}$, using the following loss function Eq.(5):

$$
\begin{aligned}
\beta^*(D) = & \, arg\min_{\beta} \lambda ||\beta||^2 \\
& + \sum_{i=1}^{N} max\left(0, 1 - y_i f_\beta(x_i)\right),
\end{aligned}
\tag{5}
$$

where $max\left(0,1-y_i f_\beta(x_i)\right)$ is the standard hinge loss and the constant λ controls the relative weight of regularization term.

DPM has been shown to provide better comic character detection performance than conventional HOG in Ref.4). However, the model of DPM proposed by Felzenszwalb et al. had an original structure that was applied to human shape detection and was not modified for comic character detection[11].

3. Features of Comic Images

General object detection can be classified into two groups. One is based on handcrafted image features and classifiers, such as SVM. The other is based on convolutional neural network (CNN) in which image feature extraction is automatically operated through training. In this study, we focus on the former approach, which uses specified image features.

Object detection using natural image features is performed in five steps: (1) Obtain the luminance information of an image, (2) Determine the feature descriptor, (3) Construct the identifier via statistical analysis, (4) Conduct the training, and (5) Detect the targeted object in an image.

Comic images have entirely different characteristics from natural images. Therefore, it is important to find a suitable feature descriptor when the detection target is in the comic image. Comic images consist of the following three components: (1) line drawings given by binary images, (2) dots printed or black filled texture area, and (3) words in speech balloons and onomatopoeia.

As a result, comic images have edge components where their intensities dynamically change and flat areas where the luminance variance is very small. Therefore, as explained in the next section, HOG features might be effective for comics because it can capture edge information.

4. Multi-View Face Detection Using DPM

Orozco et al. proposed a multi-view face detection method for human face images using DPM. In this method, they classified faces into several orientations according to their angles and allowed multi-view face detection by training root filters corresponding to each orientation. In their study, Orozco concluded that a detection model with 4 root filters and 6 part filters was optimal for face detection in natural images. However, it is likely not effective to apply this method as is to comic images. The reasons for this are, first, comic images do not have clear information for the face angle because they are two-dimensional representations and second, the feature changing of face parts in comic images is larger than that in natural images.

5. Proposed Method

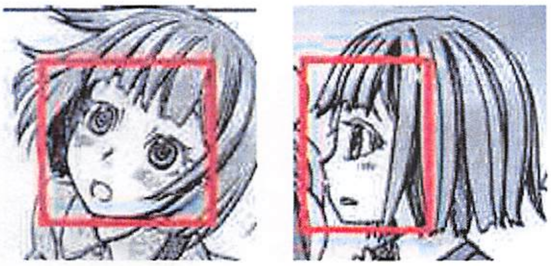
In this study, we propose a multi-view face detection method for comic characters by improving the conventional method proposed by Orozco et al. In this method, we target the faces drawn in the range from the front until 90 degrees. Firstly, in order to solve the problem that clear angle information does not exist in comic images, we classify character faces by aspect ratio instead of angles. This is based on the idea that when the angle of face changed, the aspect ratio of face region also changed. Next, we address the problem that the image feature of character face is different from that of human, by changing the number of root filters and part filters. Therefore, we determine the optimal structure of the detection model for comic images by comparing detection accuracy of DPM models with different number of root filters and part filters.

6. Character Face Detection Experiments

In this experiment, we used voc-release5 for the DPM algorithm[5]. Positive samples and negative samples for training and evaluation were set as follows. Positive samples used images cut out of the face area of comic images and obtained the annotations of the face positions. We set the face images so that both eyes were shown in front view faces and only one eye was shown in side view faces. Negative samples were images cut from regions not including face areas from comic images. Example images of positive samples and negative samples are shown in **Fig.1** and **Fig.2**. In this experiment, we considered the face detector intended for general comic works. We used the following datasets to train the detectors. The positive samples included 2000 images, 100 images each of front and side view faces from 10 comic titles. The negative samples included 2000 images randomly extracted from the same 10 comics. In addition, to evaluate the detectors, we used the following datasets. The positive samples were 1000 images that were different from those used for the training, 50 images each of front and side view faces extracted from the 10 comic titles. The negative samples were 2000 images randomly extracted from the same 10 comics. Included in the dataset used in the evaluation were positive samples that were partially occluded. Part of comic images used in this experiment, were provided by Manga 109 dataset[15].

6.1 Comparison of DPM and HOG

We compared the detection rate of DPM and HOG targeted for front and side view comic faces. In this section, we



(a) Front view face (b) Side view face

Fig.1 Example images of positive samples

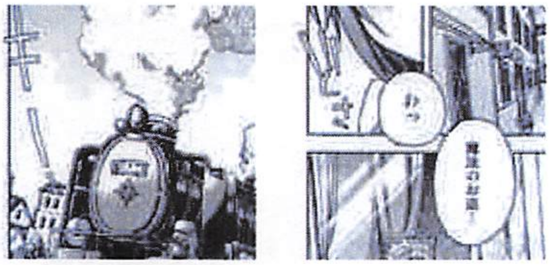


Fig.2 Example images of negative samples

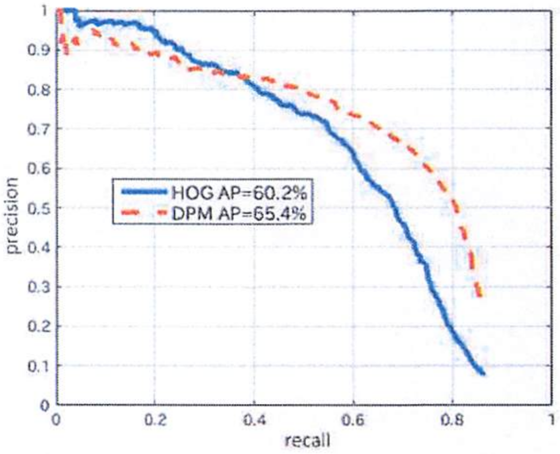


Fig.3 The Precision-Recall curves for the DPM and HOG features

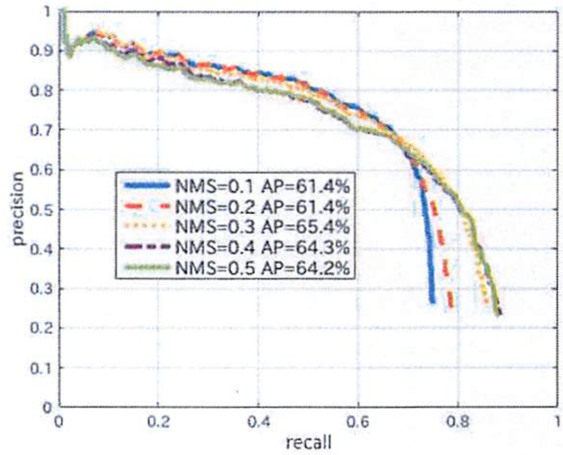


Fig.4 The Precision-Recall curves for the NMS overlap thresholds

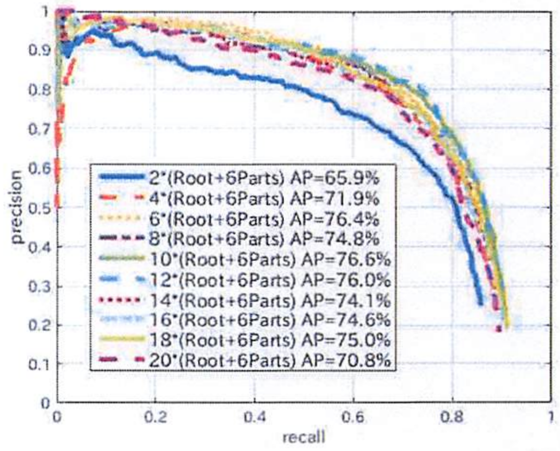


Fig.5 The Precision-Recall curves for the numbers of root filters

defined the HOG detector as the DPM detection model using only the root filters. The number of root filters was set to 2, which corresponded to the right and left sides of the character faces. The number of part filters was set to 6. The non-maximum suppression (NMS) overlap threshold was set to 0.3. The Precision-Recall curves for DPM and HOG are shown in **Fig.3**. From this figure, we see that DPM outperforms HOG by 5.2 % on average.

6.2 Optimization of NMS

Next, we describe the optimal configuration of DPM for character face detection. The NMS overlap threshold used in testing is one of the crucial parameters for the detection rate.

The evaluation result using different thresholds is shown in **Fig.4**. When the threshold value is set to 0.3, both the precision and recall rate are high and the highest average precision is obtained.

6.3 Optimization of the number of root filters

DPM can classify positive samples into *n* components by aspect ratio and train 2*n* root filters, which correspond to the right and left side of the objects. In this section, we compare the detection rate of face detection models with 2–20 root filters. We set the NMS threshold value to 0.3 and the number of part filters to 6. **Figure 5** shows a comparison of the Precision-Recall curves for the detection models with different numbers of root filters. From this figure, we see that the highest average precision value, 76.6 %, is obtained when 10 root filters are used.

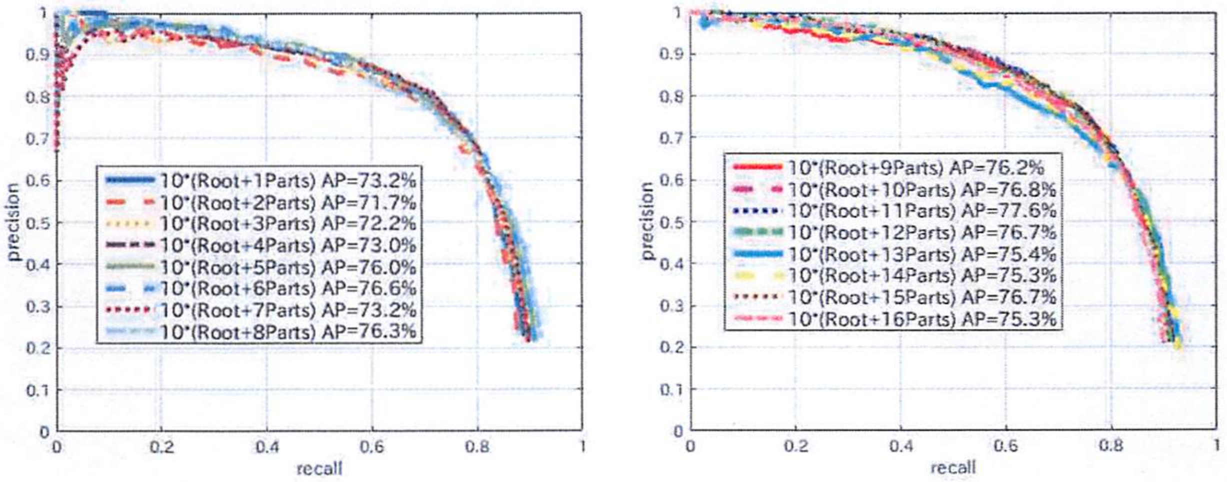
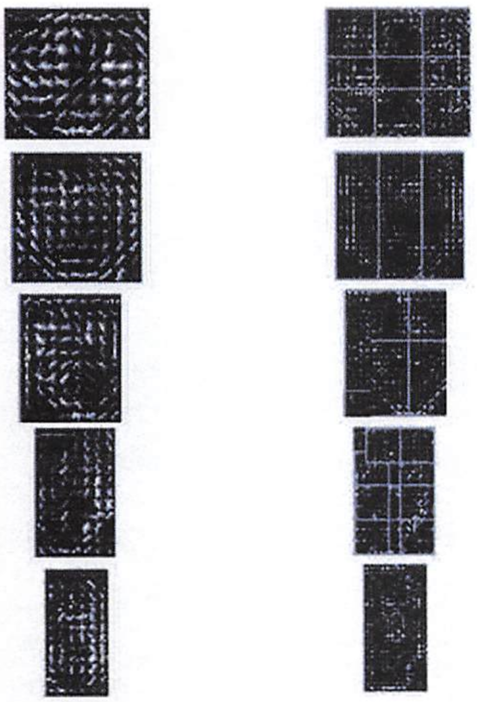


Fig.6 The Precision-Recall curves for the number of part filters



(a) Root filters (b) Part filters

Fig.7 The face detection model for comic characters

6.4 Optimization of the number of part filters

Finally, we optimized the number of part filters of DPM for character faces. The number of part filters depends on the complexity of the character faces, which are unique to the design of the comic artists. We compared the detection rate of the DPM model for 1–16 part filters. We set the NMS threshold to 0.3 and the number of root filters to 10. **Figure 6** shows a comparison of the Precision-Recall curves for the detection model with different numbers of part filters. We see that the highest average precision value, 77.6 %, is obtained when 10 root filters and 11 part filters are used. The generated detection model is shown in **Fig.7**. Figure 7(a) shows 5 components of root filters and Fig.7 (b) shows the part filters corresponding to each root filter.

7. Discussion

From the result of the experiment in Section 6.1, we see that DPM outperforms HOG in the Precision-Recall curve. In this experiment, a 5.4 % precision improvement in the average precision can be obtained. This result supports our previous results and indicates the effectiveness of the part model for comic images[4].

From the experimental results in Sections 6.3 and 6.4, we see that a detection model with 10 root filters and 11 part filters was optimal for character face detection from the comic images. However, Orozco et al. reported that a detection model with 4 root filters and 6 part filters was optimal for actual human faces[5]. Orozco et al. stated that when the number of filters increases, the precision rate tends to increase and the recall rate tends to decrease. Comic characters often have significantly different aspect ratios of faces or some of the face parts are omitted for each character. For this reason, the precision rate of character detection is lower in a face detector intended for unspecified characters than in a human face detector. Therefore, a detection model with a larger number of filters than for human face detectors is valid for comic characters.

8. Conclusion

In this study, the performance of character face detection using DPM was investigated. DPM showed higher detection in the Precision-Recall curve than HOG in our experiments. Furthermore, side view samples for character faces were included in the training and evaluation sequences. DPMs using multiple root filters were found to be suitable in this case. We found that a combination of 10 root filters and 11 part filters provided the best performance. These parameters are different from those in the case of human detection because of the particularity of comic character faces.

Acknowledgement

The authors would like to thank the anonymous reviewers for their valuable comments. The authors would also like to thank Ms. Hinata Kino (http://www.etheric-f.com/), who kindly provided all comic images used in Figure 1 and Figure 2 in this paper. This work was supported by JSPS KAKENHI Grant Number 24330137.

References

1) Internet Media Research Institute, E-comic Marketing Report 2012, Impress R&D, p.14 (2012). (in Japanese)

2) D. Ishii, H. Watanabe: "A Study on Automatic Character Detection and Recognition from Comics", The Journal of the Institute of Image Electronics Engineers of Japan, Vol.42, No.4, pp.457–465 (2013).

3) T. Arai, Y. Matsui, K. Aizawa: "Face Detection from Manga Pages", Institute of Electronics, Information, and Communication Engineers, p.161 (2012).

4) H. Yanagisawa, D. Ishii, H. Watanabe: "Face Detection for Comic Images with Deformable Part Model", Proc. of The 4th International Workshop on Image Electronics and Visual Computing (2014).

5) J. Orozco, B. Martinez, M. Pantic: "Empirical Analysis of Cascade Deformable Models for Multi-View Face Detection", Image and Vision Computing, Vol.42, pp.47–61 (2015).

6) D. Ishii, K. Kawamura, H. Watanabe: "A Study on Frame Decomposition of Comic Images", Forum on Information Technology, Vol.5, No.3, pp.263–264 (2006).

7) D. Ishii, K. Kawamura, H. Watanabe: "A Study on Frame Decomposition of Comic Images", IEICE Trans. on Information & Systems, Vol.J90-D, No.7, pp.1667–1670 (2007).

8) T. Tanaka., K. Shoji, F. Toyama, J. Miyamichi: "Layout Analysis of Tree-Structured Scene Frames in Comic Images", Proc. of 20th International Joint Conference on Artificial Intelligence, pp.2885–2890 (2007).

9) S. Nonaka, T. Sawano, N. Haneda: "Development of "GT-Scan," the Technology for Automatic Detection of Frames in Scanned Comic", Fuji - Film Research & Development, No.57, pp.46–49 (2012).

10) T. Tanaka, F. Toyama, J. Miyamichi, K. Shoji: "Detection and Classification of Speech Balloons in Comic Images", Journal of the Institute of Image Information and Television Engineers, Vol.64, No.12, pp.1933–1939 (2010).

11) P. Felzenszwalb, R. Girshick, D. McAllester, Discriminatively Trained Deformable Part Models Version 5, http://people.cs.uchicago.edu /~rbg/latent-release5/ (2012).

12) N. Dalal, B. Triggs: "Histograms of Oriented Gradients for Human Detection", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.886–893 (2005).

13) P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan: "Object Detection with Discriminatively Trained Part Based Models", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.32, No.9, pp.1627–1645 (2010).

14) P. Felzenszwalb, D. McAllester, D. Ramanan: "A Discriminatively Trained, Multiscale, Deformable Part Model", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2008).

15) Y. Matsui, K. Ito, Y. Aramaki, T. Yamasaki, K. Aizawa: "Sketch-Based Manga Retrieval using Manga109 Dataset", arXiv:1510.04389 (2015).

(Received April. 27, 2015)
(Revised April. 28, 2016)

Hideaki YAGISAWA

He received his B.S. degree from Waseda University, Japan in 2014. He is currently a M.S. student at the Graduate School of Fundamental Science and Engineering, Waseda University, Japan. His research interests are image recognition and metadata extraction.

Daisuke ISHII *(Member)*

He received his B.S. degree from the Tokyo University of Science, Japan in 2006 and his M.S. and D.S. degrees from Waseda University, Japan in 2008 and 2014, respectively. From 2009–2015 he worked as a Research Associate at Waseda University. In 2015, he moved to Fujitsu Laboratories Ltd. His research interests are image/media processing, image/media recognition, and metadata extraction. He is a member of IEICE, IIEEJ, and IPSJ.

Hiroshi WATANABE *(Member)*

He received his B.E., M.E., and Ph.D. degrees from Hokkaido University, Japan in 1980, 1982, and 1985, respectively. He joined Nippon Telegraph and Telephone Corporation (NTT) in 1985 and was engaged in research and development of image and video coding system at NTT Human Interface Labs. and Cyber Space Labs. until 2000. In 2000, he moved to Waseda University, where he is currently a Professor in the Department of Communications and Computer Engineering at the School of Fundamental Science and Engineering. His research interests include image and video processing, object recognition, and multimedia distribution. He is a member of IEEE, IEICE, ITE, IIEEJ, and IPSJ.