Detection of Facial Parts via Deformable Part Model Using Part Annotation

Kazuhiro Nishida*, Naoko Enami† and Yasuo Ariki†
* Graduate School of System Informatics, Kobe University, Japan
E-mail: nishida@me.cs.scitec.kobe-u.ac.jp
† Organization of Advanced Science and Technology, Kobe University, Japan
E-mail: naoko.enami@port.kobe-u.ac.jp, ariki@kobe-u.ac.jp

***Abstract*—In this paper, we propose a novel method for facial parts detection based on Deformable Part Model (DPM). In DPM, the parts are useful regions to detect the face and do not always correspond to the facial parts such as eye, nose and mouth. We model facial parts as a part filter and use annotation to training the position and size. In addition, we discuss the algorithm to deal with the variation of bounding box in the annotation. Our experimental results show that the proposed algorithm improves DPM for facial parts detection.**

I. Introduction

It is thought that human internal state can be estimated from eye gaze and facial expression. Currently, eye gaze estimation using iris region[1] and micro-expression recognition using strain pattern of facial regions[2] are proposed. In these approaches, it is necessary to detect facial parts such as eye, nose and mouth. The purpose of this paper is to detect facial parts accurately.

We use Deformable Part Model (DPM)[3] to detect facial parts by bounding box. The conventional DPM consists of a root filter that approximately covers an holistic object, part filters that cover smaller parts of the object, and spatial models that define a set of allowed placements of the parts relative to the root. By applying DPM to face, it can detect a face and the parts. However, the parts indicate useful regions to detect the face and do not always correspond to the facial parts such as eye, nose and mouth. To solve the problem and accurately detect the facial parts, we propose a new method to constrain part filters to locate at the position where human annotated on training images. Owing to the annotation, a face and the facial parts are detected accurately at the same time. As the annotation we use features such as size and location of facial parts and investigate the effective algorithm for facial parts detection in this paper. Fig. 1 shows the examples of our method: (a) represents a training image and bounding boxes used for annotation given by human, (b) represents part filters of DPM trained by our method and (c) represents a result of face and facial parts detection.

In addition, there is a case where bounding boxes used for annotation vary on the training images. This causes the problem that the part filters are not well decided. Hence, we discuss the algorithm to deal with the variation of bounding box used for annotation.

(a) (b) (c)

Fig. 1. Examples of our facial parts detection. (a) represents one of training images and bounding boxes used for annotation given by human. (b) represents part filters of DPM trained by proposed method. (c) represents a result of face and facial parts detection using DPM traind by proposed method.

II. Related Work

Many conventional methods in facial parts detection detect a whole face at first and then facial parts are detected in the detected facial region[4], [5]. Therefore, the precision of face detection influences the precision of facial parts detection. In addition, the conventional methods require the construction of a classifier for each facial part. On the other hand, our method can detect a face and facial parts at the same time and is not necessary to construct a classifier for each facial part. Similar to our work, part annotation is used in the works which deal with parts[6], [7], [8], [9], [10], [11]. The works [6], [7], [8] estimate human pose, [9] classifies dog and cat. Similar to our work, [10], [11] use part-based model and features such as size and location. [10] uses minimum spanning tree model. [11] uses a fully connected model where the nodes represent the holistic object and parts. In contrast, we use a star model. [10], [11] don't investigate effective algorithm in parts detection and influence of the variation of bounding box used for annotation, but we do.

III. DPM for Detection of Facial Parts

The detection score of DPM is defined as follows.

$$score = \sum_{i=0}^{n} F_i \cdot \phi(H, p_i) - \sum_{i=1}^{n} d_i \cdot \phi_d(dx_i, dy_i) + b \quad (1)$$

where F_0 is a root filter, F_i is a filter for the i-th part, $p_i = (x_i, y_i, l_i)$ specifies the feature pyramid level and position of i-th part, $\phi(H, p_i)$ denotes the feature vector in

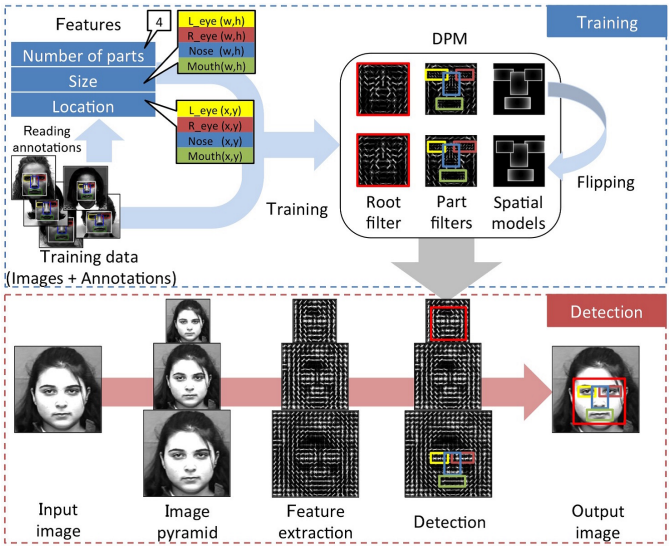


Fig. 2. Overview of the proposed method

the subwindow of feature pyramid H with top-left corner at p_i, d_i is a four-dimensional vector defining deformation cost, v_i is a two-dimensional vector specifying an anchor position for part i relative to the root position, $(dx_i, dy_i) = (x_i, y_i) - (2(x_0, y_0) + v_i)$ gives the displacement of the i-th part relative to its anchor position, $\phi_d(dx, dy) = (dx, dy, dx^2, dy^2)$ are deformation costs and b is bias term.

An overview of our approach is shown in Fig. 2. Dataset of DPM consists of images and annotations specifying a bounding box and a class label for holistic face. In this paper, we additionally annotate on facial parts of left eye, right eye, nose and mouth by using the size and location of the parts.

In DPM v5[12], the number of part filters is 8 and each part filter size is fixed (6×6). In addition, the part filters are placed to cover high-energy regions of root filter.

To detect the facial parts by part filters using annotation in our method, the parts annotation is read at first. The number of the part filters is set to the number of classes used for the parts annotation in our method. Let N be the number of training samples, $(S^x_{i,class}, S^y_{i,class})$ and $(P^x_{i,class}, P^y_{i,class})$ be the size of bounding box and top-left corner position of the bounding box annotated for *class* in i-th sample. We show a method to utilize the features (size, location) below. There are one rule in using size and two rules in using location.

- Size Rule

Size of each part filter $(size^x_{part}, size^y_{part})$ given $part \in \{left\ eye, right\ eye, nose, mouth\}$ constrained to the size of annotated facial parts as follows.

$$size^x_{part} = \lceil \frac{1}{N} \sum_{i=1}^{N} \frac{S^x_{i,part}}{S^x_{i,face}} \times 2 \times rootsize^x \rceil \quad (2)$$

$$size^y_{part} = \lceil \frac{1}{N} \sum_{i=1}^{N} \frac{S^y_{i,part}}{S^y_{i,face}} \times 2 \times rootsize^y \rceil \quad (3)$$

where the $(rootsize^x, rootsize^y)$ is the size of root filter. We call this *size rule*.

- Limitation Rule

Top-left corner position of one part filter relative to top-left corner position of a root filter is called anchor position. Anchor position of each part filter is set to cover high-energy regions within from $(P\_min^x_{part}, P\_min^y_{part})$ to $(P\_max^x_{part}, P\_max^y_{part})$ of a root filter given $part \in \{left\ eye, right\ eye, nose, mouth\}$. $P\_min^x_{part}, P\_min^y_{part}, P\_max^x_{part}$ and $P\_max^y_{part}$ are defined as follows,

$$P\_min^x_{part} = \min_{i \in \{1..N\}} \lfloor \frac{P^x_{i,part} - P^x_{i,face}}{S^x_{i,face}} \times 2 \times rootsize^x + 0.5 \rfloor \quad (4)$$

$$P\_min^y_{part} = \min_{i \in \{1..N\}} \lfloor \frac{P^y_{i,part} - P^y_{i,face}}{S^y_{i,face}} \times 2 \times rootsize^y + 0.5 \rfloor \quad (5)$$

$$P\_max^x_{part} = \max_{i \in \{1..N\}} \lfloor \frac{P^x_{i,part} - P^x_{i,face}}{S^x_{i,face}} \times 2 \times rootsize^x + 0.5 \rfloor \quad (6)$$

$$P\_max^y_{part} = \max_{i \in \{1..N\}} \lfloor \frac{P^y_{i,part} - P^y_{i,face}}{S^y_{i,face}} \times 2 \times rootsize^y + 0.5 \rfloor \quad (7)$$

We call this *limitation rule*.

- Average Position Rule

Anchor position of each part filter $(P\_anc^x_{part}, P\_anc^y_{part})$ is constrained to the averaged position of top-left corner position annotated on each facial part relative to that of the face region, given $part \in \{left\ eye, right\ eye, nose, mouth\}$. The $P\_anc^x_{part}$ and $P\_anc^y_{part}$ are defined as follows.

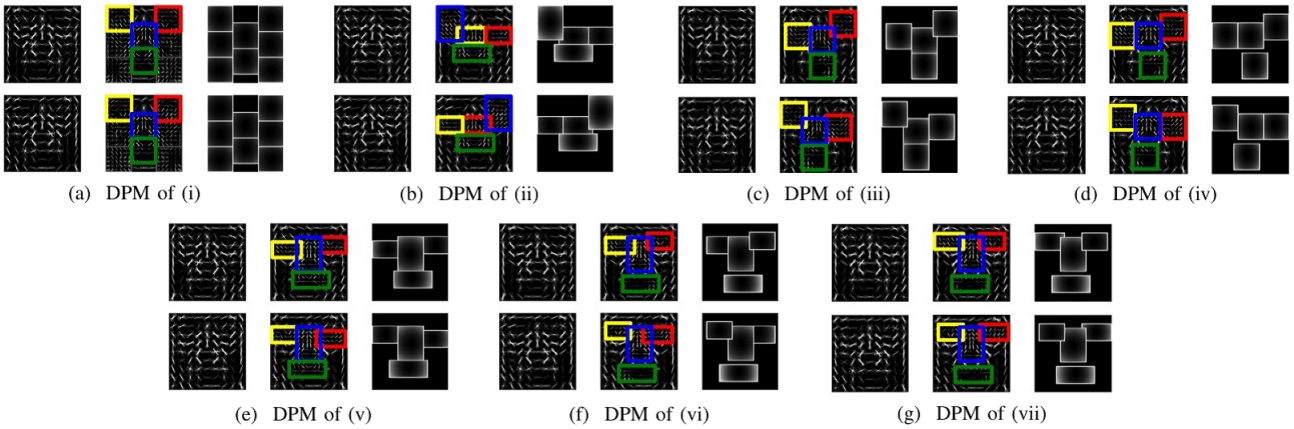
$$P\_anc^x_{part} = \lfloor \frac{1}{N} \sum_{i=1}^{N} \frac{P^x_{i,part} - P^x_{i,face}}{S^x_{i,face}} \times 2 \times rootsize^x + 0.5 \rfloor \quad (8)$$

$$P\_anc^y_{part} = \lfloor \frac{1}{N} \sum_{i=1}^{N} \frac{P^y_{i,part} - P^y_{i,face}}{S^y_{i,face}} \times 2 \times rootsize^y + 0.5 \rfloor \quad (9)$$

We call this *average position rule*.

IV. Experiments

The Extended Cohn-Kanade Dataset(CK+)[13] was used for evaluation. CK+ contains 593 sequences from 123 subjects who are 18 to 50 years old. Subjects were instructed to perform a series of 23 facial displays, six of which were based on description of prototypic emotions. For evaluation, we used 137 images of 6 subjects as the training data to train model and 200 images of 20 subjects as test data. The



(a) DPM of (i) (b) DPM of (ii) (c) DPM of (iii) (d) DPM of (iv)

(e) DPM of (v) (f) DPM of (vi) (g) DPM of (vii)

Fig. 3. DPMs trained by each method. In each model, top is a original model and bottom is the flipped model. The model consists of a root filter(left), part filters(center) and spatial models(right). In part filters, Yellow area is left eye, red area is right eye, blue area is nose and green area is mouth.

subjects are not duplicated in the training images and the test images. The bounding boxes of each facial part were given by the one person. Evaluation criterion is based on PASCAL Visual Object Classes Challenge[14]. In detection task, IoU (Intersection over Union) is calculated for each predicted bounding box B_p as follows.

$$\mathrm{IoU}(B_p, B_{gt}) = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{10}$$

where B_{gt} is a ground truth bounding box. A correct detection is decided when IoU exceeds overlap threshold t_{ov}. As t_{ov} are used $0.5, 0.6, 0.7, 0.8, 0.9$ in our face detection, and $0.3, 0.4, 0.5$ in our facial parts detection. Average Precision (AP) is used as the evaluation of the detections.

We compared 6 methods to show the effectiveness of the rules used in facial parts detection. They are (i) DPM v5 and proposed methods using (ii) *size rule*, (iii) *limitation rule*, (iv) *average position rule*, (v) *size rule* + *limitation rule* and (vi) *size rule* + *average position rule*. Part filters of (ii)-(vi) correspond to each facial part. For (i), we evaluated AP of facial parts detection by computing the detected results with the annotation. As a result, 4 part filters corresponding to each facial part of left eye, right eye, nose and mouth were selected from 8 part filters. The constructed DPMs (containing flipped horizontal model) trained by each method are shown Fig. 3(a)-(f). In these figures, the part filters corresponding to left eye, right eye, nose and mouth are shown in yellow, red, blue and green respectively.

Result of facial parts detection by each method is shown in Table I. Then, mean of mean AP by each method is shown in Table II. Method (vii) will be described later. As for face detection, mean APs (%) are (i) 69.6, (ii) 67.2, (iii) 69.1, (iv) 67.9, (v) 64.2, (vi) 67.3. There are no significant difference between (i) DPM v5 and (ii)-(vi) proposed methods. The detection time is shown in Table III using Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz.

V. Discussion

Method (ii) using *size rule* only shows the worst result. Therefore, importance of *limitation rule* and *average position rule* were verified to be effective. Results by method (v)(vi) using *size rule* plus *limitation rule* and *average position rule* are better than method (iii)(iv) using *limitation rule* only and *average position rule* only. Thus, *size rule* is verified to be effective. By comparison of (v) and (vi), *average position rule* is more important than *limitation rule*. Furthermore, since (vi) shows the best among the comparative methods, the effectiveness of combining *size rule* and *average position rule* in facial parts detection is verified to be effective.

Results of right eye and mouth detection are relatively worse. We assume this problem is due to the variation of the bounding boxes in the annotations on training images. To solve this problem, we propose to change the *average position rule* to calculate the difference between center of gravity positions of bounding box of face and each facial part, instead of calculating the difference between top-left corner positions of them.

Anchor position of each part filter ($P\_anc^x_{part}$, $P\_anc^y_{part}$) given $part \in \{left\ eye, right\ eye, nose, mouth\}$ is defined as follows,

$$P\_anc^x_{part} = \lfloor \frac{1}{N} \sum_{i=1}^{N} \frac{P^{\mu_x}_{i,part} - P^{\mu_x}_{i,face}}{S^x_{i,face}} \times 2 \times rootsize^x + rootsize^x - \frac{size^x_{part}}{2} + 0.5 \rfloor \tag{11}$$

$$P\_anc^y_{part} = \lfloor \frac{1}{N} \sum_{i=1}^{N} \frac{P^{\mu_y}_{i,part} - P^{\mu_y}_{i,face}}{S^y_{i,face}} \times 2 \times rootsize^y + rootsize^y - \frac{size^y_{part}}{2} + 0.5 \rfloor \tag{12}$$

TABLE I
AVERAGE PRECISION (%) OF FACIAL PARTS DETECTION

	left eye				right eye				nose				mouth			
	overlap threshold t_{ov}				overlap threshold t_{ov}				overlap threshold t_{ov}				overlap threshold t_{ov}			
	0.3	0.4	0.5	mean AP	0.3	0.4	0.5	mean AP	0.3	0.4	0.5	mean AP	0.3	0.4	0.5	mean AP
(i)	68.6	18.3	0.5	29.1	33.2	6.6	0.2	13.3	**98.3**	**98.3**	**97.8**	**98.1**	88.5	35.9	0.5	41.6
(ii)	0.3	0.1	0	0.1	0.6	0.4	0	0.3	0.1	0	0	0.3	0.2	0.1	0	0.1
(iii)	40.8	5.8	0.7	15.8	7.3	0.7	0.2	2.7	96.1	90.6	58.3	81.7	84.2	28.9	0.1	37.7
(iv)	25	4	0.2	9.7	6.4	2.1	0.2	2.9	96.7	96.7	96.7	96.7	83.9	30.1	0.4	38.1
(v)	60	38	19.8	39.2	8	6.5	3.7	6.1	96.6	96.6	90.3	94.5	51.9	9.2	0.8	20.6
(vi)	**96.7**	84.4	57.7	79.6	**96.7**	**64.3**	21.4	60.8	96.7	96.7	88.2	93.9	92.9	**84.3**	58.7	78.6
(vii)	96.1	**85.9**	**61**	**81**	88.7	60.2	**36.4**	**61.8**	96.1	96.1	89.4	93.9	**96.1**	84.1	**77**	**85.7**

TABLE II
MEAN OF MEAN AP (%) IN FACIAL PARTS DETECTION

(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
45.6	0.15	34.5	36.9	40.1	78.2	**80.6**

TABLE III
DETECTION TIME (SECOND) PER IMAGE

(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
1.569	2.891	2.924	2.918	2.907	2.902	2.909

where $P^{\mu_x}_{i,face} = P^x_{i,face} + rootsize^x/2$, $P^{\mu_y}_{i,face} = P^y_{i,face} + rootsize^y/2$, $P^{\mu_x}_{i,part} = P^x_{i,part} + size^x_{part}/2$, $P^{\mu_y}_{i,part} = P^y_{i,part} + size_part^y/2$. (vii) is a method using this rule and *size rule*. DPM trained by (vii) is shown in Fig. 3(g). The detection example of (vii) is shown in Fig. 1(c).

There are annotation with tight bounding box as well as loose bounding box in surrounding one object in training image. Therefore, it is assumed that center of gravity position has smaller variation than top-left corner position.

The problem of annotated bounding box variation is not only in the case of face and facial parts but also in others. In particular, the variation causes negative impact on calculating the relations between two objects as is in this case. we did annotation by one person in this experiment. In the case of annotation by many persons, the variation will be larger. Thus, it is important to investigate the method to reduce the variation.

Furthermore, in method (vii), we also evaluate small, occluded and motion blurred faces because assuming real environments. Small image is half size (320×245 pixels) of the original image (640×490 pixels). Size of occluded region is 10%, 20% and 30% of face size and number of the regions is 1 or 2. Occluded regions are chosen by random from four facial parts. Number of occluded images is 1200. As for motion blur, the length of camera motion is 20 pixels and the angle θ is 0°, 45°, 90° and 135°. Number of motion blurred images is 800.

Results of facial parts detection (mean of mean AP (%)) are 82.6(small), 40.8(occluded) and 74.9(motion blurred).

VI. CONCLUSION

In this paper, we proposed a new method to detect accurately facial parts such as eye, nose and mouth using DPM. In order to constrain part filters to locate at the position where human annotated on training images, we investigated the effective algorithm using size and location of bounding box used in the annotation. In addition, we discussed the algorithm to deal with the variation of bounding box in the annotation. As a result, the proposed algorithm was verified to be effective in facial parts detection, which can constrain the size of part filter by the size of annotated facial parts as well as the anchor position of each part filter by the averaged position of annotated facial parts position. To reduce the variation of bounding box used in the proposed annotation, the proposed algorithm was verified to be effective, which can decide the anchor position of each part filter by calculating the difference between center of gravity positions of bounding box of face and each facial part.

REFERENCES

[1] W. Waizenegger, N. Atzpadin, O. Schreer, I. Feldmann, P. Eisert, Model based 3D gaze estimation for provision of virtual eye contact, Proc. ICIP, 2012.

[2] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, S. Sarkar, Towards macro- and micro-expression spotting in video using strain patterns. In Workshop on Applications of Computer Vision, pages 1-6, 2010.

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9, Sep. 2010.

[4] K. Peng, L. Che, S. Ruan, and G. Kukharev, A Robust Algorithm for Eye Detection on Gray Intensity Face without Spectacles, J. Computer Science & Technology, vol.5, no.3, pp.127-132, 2005.

[5] L. Nanni, A. Lumini, Combining face and eye detectors in a high-performance face-detection system, IEEE MultiMedia, vol. 19, No. 4, pp. 20–27, Oct.–Dec. 2012.

[6] Y. Yang and D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, In CVPR, 2011.

[7] Y. Wang, D. Tran, and Z. Liao, Learning hierarchical poselets for human parsing, In CVPR, 2011.

[8] S. Johnson and M. Everingham, Learning Effective Human Pose Estimation from Inaccurate Annotation, In CVPR, 2011.

[9] O. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman, The truth about cats and dogs, In ICCV, 2011.

[10] H. Azizpour and I. Laptev, Object detection using strongly-supervised deformable part models, In ECCV, 2012.

[11] X. Chen, R. MottaghChen2i, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, Detect what you can: Detecting and representing objects using holistic models and body parts, In CVPR, 2014.

[12] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, Discriminatively trained deformable part models, release 5. ⟨http://www.cs.berkeley.edu/~rbg/latent/⟩

[13] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in Proceedings of the IEEE Workshop on CVPR for Human Communicative Behavior Analysis, 2010.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, The PASCAL Visual Object Classes Challenge, International Journal of Computer Vision (IJCV), 88(2):303–338, June 2010.