PAPER

Iterative Improvement of Human Pose Classification Using Guide Ontology

Kazuhiro TASHIRO[†], **Takahiro KAWAMURA**[†,††a)], *Nonmembers*, **Yuichi SEI**[†], **Hiroyuki NAKAGAWA**[†††], *Members*, **Yasuyuki TAHARA**[†], *Nonmember, and* **Akihiko OHSUGA**[†], *Member*

SUMMARY The objective of this paper is to recognize and classify the poses of idols in still images on the web. The poses found in Japanese idol photos are often complicated and their classification is highly challenging. Although advances in computer vision research have made huge contributions to image recognition, it is not enough to estimate human poses accurately. We thus propose a method that refines result of human pose estimation by Pose Guide Ontology (PGO) and a set of energy functions. PGO, which we introduce in this paper, contains useful background knowledge, such as semantic hierarchies and constraints related to the positional relationship between body parts. Energy functions compute the right positions of body parts based on knowledge of the human body. Through experiments, we also refine PGO iteratively for further improvement of classification accuracy. We demonstrate pose classification into 8 classes on a dataset containing 400 idol images on the web. Result of experiments shows the efficiency of PGO and the energy functions; the F-measure of classification is 15% higher than the non-refined results. In addition to this, we confirm the validity of the energy functions.

key words: *ontology, semantic web, knowledge representation*

1. Introduction

From the late 2000's, a large number of idol groups have appeared and gained tremendous popularity in Japan. The fiercely competitive situation with regard to Japanese idols is called "Idol Sengoku Jidai" (literally "Warring States Period of Japanese Idol"). The number of idol photographs on the web has increased explosively, but user preferences about the poses of idols may vary. Thus, there seems to be potential demand for searching for the idol photos by their poses. Our goal is to automatically classify the idol photos according to their poses during web searches. This is challenging due to uncontrolled conditions under the photos were taken with very cluttered background, for example, idols appearing at a wide range of scales in dark illumination. Thus, search engines such as Google images* and Microsoft Bing images** do not recognize the poses of people in images. Our method has two phases; the first phase is to estimate the spatial layout of ten body parts (head, torso, upper and lower arms and legs) using Eichner's Stickman Pose Estimation. The second phase is to amend estimation result based on PGO and to classify the poses of the idols using Bayesian Network classifiers. In this paper, we focused on the idols who are wearing swimsuits, since more than half of the image search results for keywords related to the idols are swimsuit images.

The remainder of this paper is organized as follows. We first review of related works on human pose estimation and ontology in Sect. 2. Then, the overview of proposed method in Sect. 3. In Sect. 4 and Sect. 5, we describe the details of proposed method including our pose estimation and pose classification mechanism. In Sect. 6, we conduct experiments to evaluate the proposed method, and refine Pose Guide Ontology (PGO) based on the results of classification in Sect. 7. Finally, we conclude this paper with a discussion and future work in Sect. 8. Although we processed and evaluated the idol photos retrieved from the web, figures in this paper are replaced with dolls due to publication restriction.

2. Related Work

For several decades, human pose estimation and classification have been studied in order to deal with specific situation. For example, in the case of in-vehicle camera [1], the pose classification aims to determine the orientation of pedestrians and their movement. Also, in the case of surveillance camera [2], it aims to detect a criminal act according to poses such as punching and kicking from movies. These studies, however, classify human poses into very few classes, and handle simple poses. On the other hands, our target poses are complicated and not in movies but still images. Also, we do not need the real-time estimation and thus applied ontology constrained technique, which may take time but high accuracy.

The main techniques proposed in this paper are human pose estimation and amendment by ontology. Following 2 subsections review the existing studies related to these techniques.

2.1 Related Work of Stickman Pose Estimation

A number of approaches to the human pose estimation have

Manuscript received March 2, 2015.
Manuscript revised August 10, 2015.
Manuscript publicized October 1, 2015.

[†]The authors are with Graduate School of Information System, The University of Electro-Communications, Chofu-shi, 182–8585 Japan.

[††]The author is with Department of Information Planning, Japan Science and Technology Agency, Tokyo, 102–8666 Japan.

[†††]The authors are with Graduate School of Information Science and Technology, Osaka University, Osaka-shi, 565–0871 Japan.

a) E-mail: kawamura@ohsuga.is.uec.ac.jp (Corresponding author)

DOI: 10.1587/transinf.2015EDP7067

*http://www.google.com/imghp

**http://www.bing.com/images

been studied in recent years. For pose estimation, various methods have been employed such as exploiting parallelism of part boundaries from segmentation images [3] and tracking body contour using color and depth cue [4]. Our study builds on Pictorial Structures Model [5], a popular framework for human pose estimation in still images. Pictorial Structures Model consists of articulated rectangular body parts connected in a tree topology that encodes relative part positions and orientations. In this framework, advancements were made by Ramanan [6], who proposed image parsing algorithm, which learns Pictorial Structures Model parameters by using iterative estimation. Based on Ramanan's parsing algorithm, moreover, Eichner et al. proposed a technique for estimating the human poses in highly challenging almost uncontrolled images, without prior knowledge of background, clothing, lighting, or the location and scale of the person in the image [7].

In Eichner's approach, six body parts (head, torso, upper arms, and lower arms) are expressed as sticks of a fixed size with position of location and orientation parameters, and thus this approach is called Stickman Pose Estimation. They estimates the spatial layout of body parts in a still image of a television drama scene. The approach assumes only the pose of a person who are standing upright, since most people in TV shows or movies appear roughly upright. Our method follows the Stickman Pose Estimation, but the idol photos do not necessarily have the upright posture. Therefore, we improved this work in order to deal with non-upright postures.

2.2 Related Work of Guide Ontology

Several approaches combining computer vision and ontology to understand images have been proposed. Marszalek and Schmid proposed to use lexical semantic networks to extend object recognition techniques [8]. They employed WordNet [9] to integrate background knowledge about hierarchies of real world objects into visual appearance of detected objects in images. Nwogu et al. studied the scene recognition of still images using ontology that contains semantic hierarchies of both object classes and relation classes [10]. Their ontology contains information such as "Scene has a sky", "Sky above land", and "land has a person". Then, it transfers these entities and their relations to their location relations in still images. However, location information is not enough to understand the real world.

We gave special attention to the scene recognition method using Guide Ontology (GO) proposed by Chen et al. [11]. GO does not contain location relations of object classes, but also it supports various semantic constraints. They employed the GO as a semantic source of background knowledge, and proposed Object Relation Network to transfer rich semantics in the GO to the detected object and their relations in the image. Figure 1 shows that the Object Relation Network represents three person nodes labeled "Soccer Player", a ball node labeled "Soccer Ball", and two relation nodes labeled "kick". These nodes are labeled temporarily based on visual features such as size, color, relative position of each object, not considering the semantics of the objects and their relations. Inappropriate Object Relation Network (e.g., Basket Ball Player Kick Soccer Ball) can be generated when temporary labeling fails. In order to amend this semantic failure, the Guide ontology is used, which contains semantic constraints (e.g., given Object Relation Network as Soccer Player Kick Ball, Ball node should be labeled Soccer Ball).

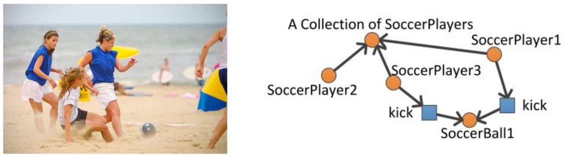


Fig. 1 Object relation network [11]

We employed PGO in order to estimate and classify human pose in still image. Both PGO and GO help to recognize still image, but their purposes are different. While GO corrects the result of labeling based on visual features after the classification, PGO corrects the estimation results of the parts location before the classification. To correct the classification results like GO, it is necessary to increase classes and properties for constraints according to the number of classified poses. On the other hand, since PGO corrects each body part and generates feature vectors for the classification. Thus, the classes and properties are limited to the number of body parts.

3. Proposed Method for Pose Classification

Overview of our method is illustrated in Fig. 2. Our method can be separated into the pose estimation and the pose classification phase. The pose estimation phase estimates the location, orientation, and size of the body parts, and the pose classification phase classifies the photos by the poses based on parts information obtained by the pose estimation. These phases are described in Sect. 4 and Sect. 5, respectively. Our method first takes the idol images, which are not annotated by any text as input, and then employs a face detector to obtain the information about the rough position and scale of a human body in the image (Sect. 4.1.1). Next, our method segments foreground parts from background parts based on Grabcut [12] and the skin color phase (Sect. 4.1.2). Then, we obtain the information of location, orientation, and the size of body parts at the parsing stage (Sect. 4.1.3). To improve the parts information, we use PGO, which contains useful background knowledge such as semantic hierarchies and constraints related to the orientation and positional relationship between the body parts (Sect. 5.2). PGO amends the location, orientation, and the size of body parts obtained by the pose estimation. We will describe PGO in Sect. 5.1. Finally, Bayesian Network classifier outputs the result of pose classification (Sect. 5.3).

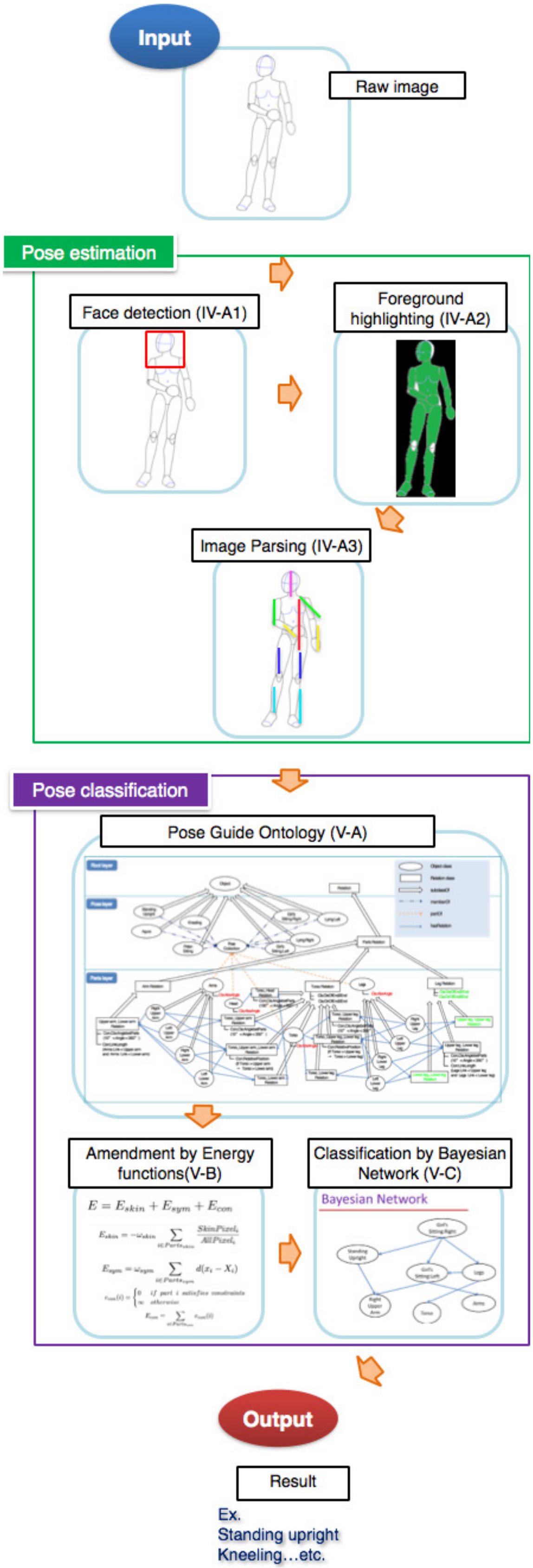

Input
Raw image
Pose estimation
Face detection (IV-A1)
Foreground highlighting (IV-A2)
Image Parsing (IV-A3)
Pose classification
Pose Guide Ontology (V-A)
Amendment by Energy functions(V-B)
$E = E_{skin} + E_{sym} + E_{con}$
Classification by Bayesian Network (V-C)
Bayesian Network
Output
Result
Ex.
Standing upright
Kneeling…etc.

Fig. 2 Proposed method for pose classification

4. Pose Estimation Phase

4.1 Approach to the Pose Estimation

4.1.1 Face Detection

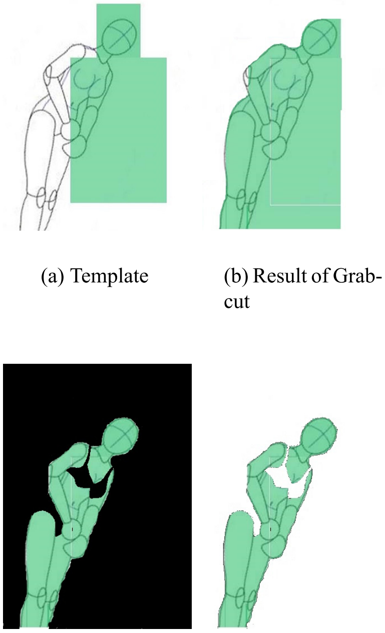
First, our method uses a face detector to find an approximate position of a person in a photo. This process removes part of the background clutter and constrains the search space in a rectangle. Eichner's method employs an upper-body detector, which has excellent accuracy on the related task of rigid object detection [13]. However, an upper-body detector cannot detect a non-upright person.

Therefore, we employed a face detector in OpenCV face detector. Our method first detects a face, more precisely a face with the neck and part of the shoulder, mainly to obtain the difference between the colors of a person and of the background. Then, by using those color information the method explores in an image and extend a rectangle to include the whole body. Especially, arms should start from the shoulder, and the torso starts from the neck and has the legs on the opposite side, so that the method can set the starting points there. Thus, slant/lying poses can be included in the rectangle.

4.1.2 Foreground Highlighting Using Grabcut and Skin Color

The pose estimation aims to localize the spatial layout of human body parts. The foreground area is likely to contain human body parts, but the background is not. Thus, the foreground highlighting helps estimating the layout of body parts. Eichner employs Grabcut algorithm [12] for highlighting the foreground. Grabcut algorithm requires the prior input and learning of initial foreground and background color models from regions, where the person is likely to be there. The foreground color model is learnt from a region, where the foreground-template (green convex region in Fig. 3 (a)) covers. But, in Eichner's work, the accuracy of foreground highlight for non-upright postures is low, since the foreground-template is specialized in the upright posture. In Fig. 3 (a), the foreground-template partially covers the background area, since a person is inclining to the right side. As a result, the background area in Fig. 3 (b) is mistakenly highlighted as the foreground area (green region is highlighted as the foreground area).

Most of the idols in photos do not have the upright posture. Therefore, the approach only using the foreground-template is not suitable for our purpose, and thus we used the hue of skin in addition to Grabcut algorithm by the template. Figure 3 (c) is a skin color detection image. Skin area is extracted using a skin color model based on the hue of pixel. Finally, we obtain the foreground highlight image (Fig. 3 (d)), which is the intersection of Fig. 3 (b) and Fig. 3 (c). This simple technique increased the accuracy of



(a) Template

(b) Result of Grabcut

(c) Skin detection

(d) Grabcut + Skin detection

Fig. 3 Foreground highlighting using Grabcut and skin color

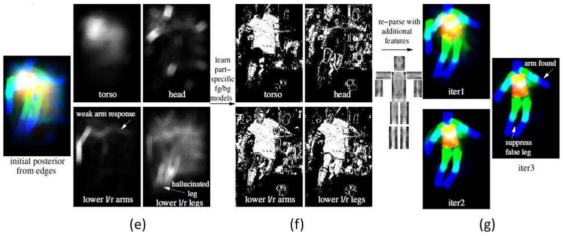


Fig. 4 Image parsing

foreground highlight[†].

4.1.3 Image Parsing

At this stage, image parsing employs Pictorial Structure Model [14], which is a general body model. In the Pictorial Structure Model, the locations of body parts are estimated based on parameters such as appearance and spatial information. In the first estimation, the model considers only image edges restricted to the foreground area. This process temporarily estimates the body part position, which are used to build appearance models of the body parts. Then, we considered image-specific appearance models based on the first estimation, and the model becomes more accurate. As a result of the estimation, the location information of ten body parts is obtained.

5. Pose Classification Phase

In this phase, the idol photos are classified based on the location information obtained in Sect. 4.1.3. Although our method employs the face detector and the foreground highlight based on the skin color information, the accuracy of location information is not reliable enough, and it is difficult to classify the poses of idols based on the incorrect location information. Therefore, we propose the approach to amend the location information of body parts using PGO.

5.1 Pose Guide Ontology

An example of PGO is shown in Fig. 5. PGO has three layers; root layer, pose layer, and parts layer. Root layer contains general classes like Object and Relation. Pose layer contains pose classes such as Standing upright, and Lying, which is output of our method. Parts layer is the most important layer, which contains semantic background knowledge about the relations between the body parts. Each object class at this layer must be a part of the pose classes in the Pose layer and each relation class must be a subclass of Parts Relation class. Parts Relation class is divided into Torso Relation, Arm Relation, and Leg Relation class. These classes have individual properties. In Fig. 5, properties are described under each class. The properties which have "Con" sign are used as constraints for amendment, and the properties which have "Cla" sign are used to generate features for classification. PGO amends the location information based on these constraints (Sect. 5.2). These properties are defined in each part class and inter-parts class, and then the properties in parent classes are effective also in child classes. For example, the Torso Relation class has the Cla:DisOfEnd2End property used for Feature 2. This property is also applicable to a child class of the Torso Relation class, the Torso_Upper leg Relation. Thus, the representation of PGO using class inheritance of ontology has high extensibility and maintainability e.g., in the case that new properties are added. Currently, PGO has been manually created. In the near future, we plan to address the automatic creation of the PGO.

The following procedure illustrates how PGO is used for the pose classification.

1. First the pose estimation is conducted.
2. Then, the estimation result is compared with constraints in properties in PGO (the properties which have "Con" sign in Fig. 5). The constraints provided by PGO are desricibed in the following subsections.
3. If parts violate the constraints, the amendment procedure is executed as described in Sect. 5.2.4.
4. According to properties for feature generation (the properties which have "Cla" sign in Fig. 5), feature vectors are generated from amended parts positions.
5. Bayesian Network classifies the poses based on the feature vectors.

[†] We tried both process flows, that is, Grabcut → Skin and Skin → Grabcut, and then selected the current flow, since it had generally high accuracy.

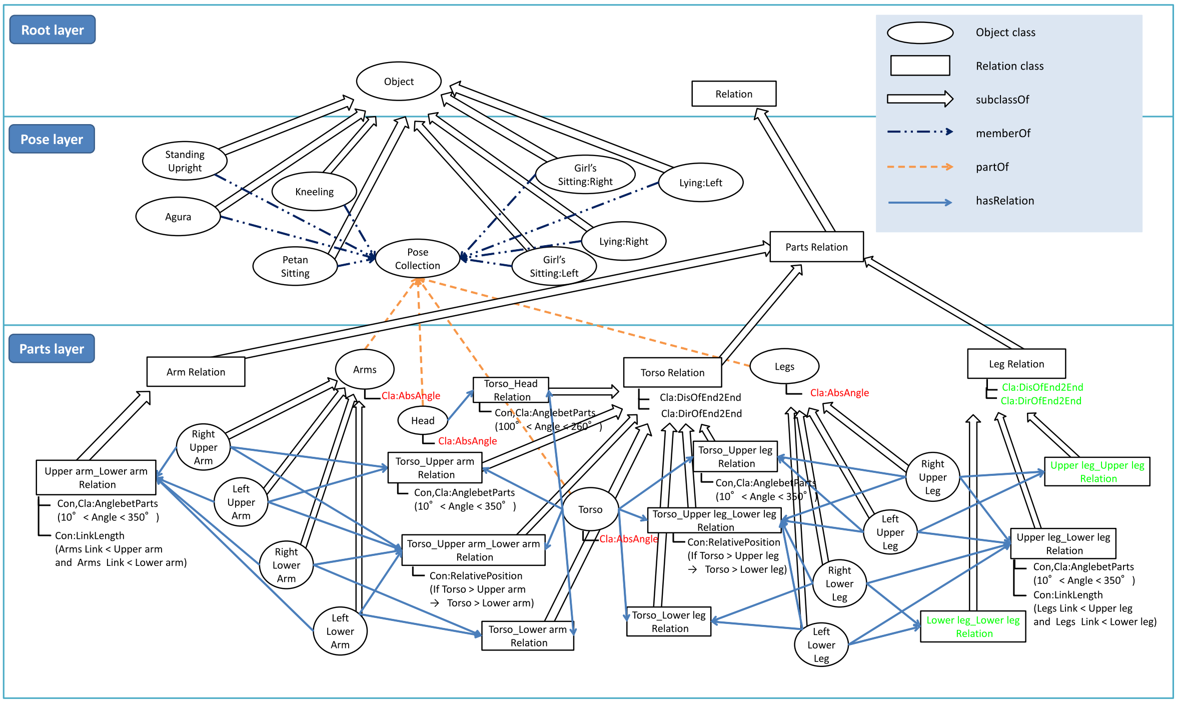


Fig. 5 An example of PGO. Each Object class and Relation class has individual properties, which are described under the classes. The ones described in red letters are added in Iteration 2(7.1), and green letters are added in Iteration 3(7.3)

5.1.1 Constraints on the Angles between the Parts

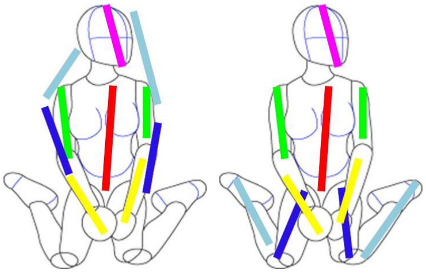
A person has an appropriate joint range of motion, which defines appropriate angles between the body parts. For example, the head and the torso are in the same direction in many cases, and it cannot be assumed that head-torso angle is 30° (torso angle defaults to 0°). Therefore, PGO has a constraint such as “head-torso angle should be in a range from 100° to 260°” (Fig. 5: Head-Torso Angle).

5.1.2 Constraints on Relative Position between the Parts

When the location of a body part is considered as inappropriate based on the relative position to other body parts, this location information should be amended. For example, PGO has constraints such as “when upper legs lie at lower than torso, lower legs also should lie in lower than torso” (Fig. 5: U.leg-L.leg-Torso Relative Position).

5.1.3 Constraints on the Links between the Parts

The link represents a joint of human body parts connecting the parent part (e.g. upper arms) and the child part (e.g. lower arms). When the location information of the body parts is estimated correctly, the lengths of links are relatively short. Therefore, PGO has a constraint such as “the length of upper legs and lower legs should be longer than one of the legs-links” (Fig. 5: U.leg-L.leg Link-Length).



(a) before amendment (b) amended

Fig. 6 **sticks color, pink**: head, **red**: torso, **green**: upper arms, **yellow**: lower arms, **blue**: upper legs, **light blue**: lower legs

5.2 Amendment by PGO and Energy Functions

A result of Stickman Pose Estimation is shown in Fig. 6: (a), and the amended result by PGO is shown in Fig. 6: (b). The constraints of PGO move the body parts from inappropriate positions to right ones. In order to compute right positions of body parts, we introduced a set of energy functions to take three kinds of knowledge into account: (I) skin color information, (II) symmetry of the human body, and (III) constraints of the PGO.

$$E = E_{skin} + E_{sym} + E_{con} \tag{1}$$

Equation (1) represents a sum of the skin color energy, the symmetry energy, and the constraint energy; which are detailed in the following subsections respectively. When the energy function E is minimized, location of the body parts is optimized.

5.2.1 Energy: Skin Color Information

The appropriate positions of body parts are expected by the skin color. Upper arms and legs, and lower arms and legs have the skin color in the case of a person wearing a swimsuit, and thus we consider this energy is important. The energy is applied to $Parts_{skin}$ = {right upper arm, left upper arm, right lower arm, left lower arm, right upper leg, left upper leg, right lower leg, left lower leg}. The skin color information based energy is defined as:

$$E_{skin} = -\omega_{skin} \sum_{i \in Parts_{skin}} \frac{SkinPixel_i}{AllPixel_i} \tag{2}$$

, where ω_{skin} is a weight, $AllPixel_i$ is the total number of pixels around candidates for the location of part i, and $SkinPixel_i$ is the number of skin-colored pixels in $AllPixel_i$.

5.2.2 Energy: Symmetry of Human Body

There are many asymmetry poses in images, but starting points of upper arms and upper legs would be mostly arranged symmetrically with a torso as a center, due to the structure of the human body. Therefore, we applied the energy for the symmetry to $Parts_{sym}$ = {right upper arm, left upper arm, right upper leg, left upper leg}. We define the symmetry of the human body based energy as:

$$E_{sym} = \omega_{sym} \sum_{i \in Parts_{sym}} d(x_i - X_i) \tag{3}$$

, where ω_{sym} is a weight. Xi and xi stand for horizontal pixel points in an image, but Xi is a tentative starting point of left upper arm, right upper arm, left upper leg, and right upper leg, arranged symmetrically with a torso as a center. Also, xi is a candidate point, from which each part starts. $d(x_i - X_i)$ is a distance between a candidate for a start point x_i and a tentative start point X_i. Thus, Eq. (3) represents that if the candidate points are arranged symmetrically, then the energy gets lower, and the points have properly estimated positions.

For example, when the left or right upper leg violates constraints in PGO, the tentative starting point Xi of the right or left leg is arranged symmetrically across the torso with another leg. If both upper legs violate the constraints, then they are arranged from the end of the torso with the fixed angle (45° on either side) and the fixed length (a quarter of the torso). These procedures correspond to third item in Sect. 5.2.4 and (2), (3) in Fig. 7. Xi is a blue point in the figure.

To evaluate this energy, we conducted experiments in Sect. 6.4.

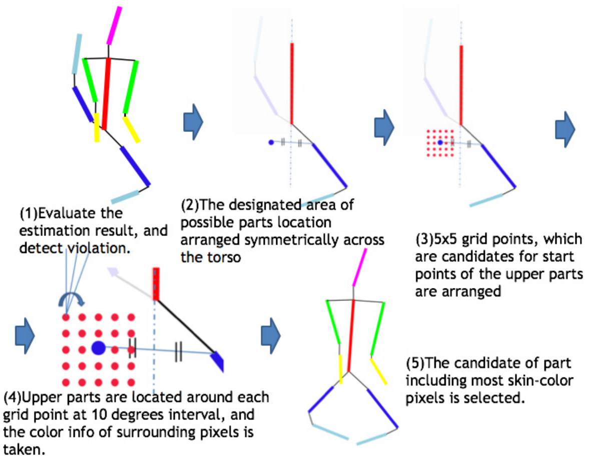

(1)Evaluate the estimation result, and detect violation.
(2)The designated area of possible parts location arranged symmetrically across the torso
(3)5x5 grid points, which are candidates for start points of the upper parts are arranged
(4)Upper parts are located around each grid point at 10 degrees interval, and the color info of surrounding pixels is taken.
(5)The candidate of part including most skin-color pixels is selected.

Fig. 7 Amendment procedure

5.2.3 Energy: Constraints of PGO

This energy is applied to all body parts, $Parts_{con}$ = {torso, right upper arm, left upper arm, right lower arm, left lower arm, right upper leg, left upper leg, right lower leg, left lower leg, head}. We define this energy function based on background knowledge in PGO.

$$e_{con}(i) = \begin{cases} 0 & if\ part\ i\ satisfies\ constraints \\ \infty & otherwise \end{cases} \tag{4}$$

$$E_{con} = \sum_{i \in Parts_{con}} e_{con}(i) \tag{5}$$

Following Chen's energy function [11], E_{con} adds a big penalty to candidates of body parts when any of the constraints is violated. Therefore, the candidates of body parts which violate the constraints are not adopted. We are considering a soft decision for the constraints of PGO in Eq. (4) or even a parameter to be learned.

5.2.4 Amendment Procedure Based on Energy Functions

Amendment procedure is detailed below with Fig. 7.

1. Evaluate the result of Stickman Pose Estimation, and detect constraints violation (Fig. 7 (1)).
2. If any limbs violated the constraints provided by PGO, amendment procedure is executed (Upper arms or legs, which do not violate, are amended only if lower arms or legs violated).
3. The designated area of possible parts location are arranged symmetrically across the torso (Fig. 7 (2)).
4. 5x5 grid points, which are candidates for start points of the upper arms or legs, are arranged in the designated area. The width and height of the area are set as the same as half length of torso (Fig. 7 (3)).
5. Upper parts are located around each grid point at 10 degrees interval, and the color information of surrounding

pixels is taken at each location (Fig. 7 (4)). For reducing calculation cost, lower 50% of parts location with relatively small skin-color pixels are removed.

6. Also, lower parts are located around each end point of upper parts, and the color information of surrounding pixels is taken at each location.
7. The candidate of part including most skin-color pixels is selected (Fig. 7 (5)).
8. If the selected parts violate the constraints provided by PGO, the second most skin-colored one is adopted.

5.3 Pose Classification Using Bayesian Network

After the location information of the body parts is amended, our method generates input data (feature vectors) for Bayesian Network by referring properties for feature generation in PGO (signed "Cla" in Fig. 5). The Bayesian Network outputs the probability of each pose based on angles between two parts and distance between the torso and other parts. In our experiments, a feature vector has 87 dimensions of numerical features that represent human body layout. We derived it from 400 training images. In this paper, we employed weka.classifiers.bayes.BayesNet, which is prepared in Weka[†]. Actual classification and feature selection are detailed in the next section.

Since we conduct iterative PGO refinements in Sect. 7, we need information about which feature contributes to the accuracy improvement, that is, the importance of each feature. Therefore, a requirement of the classifier choice was that it is easy to understand how the feature is used in the classification process. As such classifiers, there are Random Forest, Bayesian Network, and so on, but the classification process of Bayesian Network can be visible, and also its accuracy had higher than others in the following experiments.

6. Iterative Experiment

To evaluate the improvement of accuracy, we demonstrate how our method classifies the idol photos in five different scenarios for comparison. Figure 8 shows (a)~(e) scenarios and purpose of them. First, we explain two scenarios for evaluation of PGO: (a) using only Stickman Pose Estimation improved in Sect. 4.1.1 and Sect. 4.1.2, and (b) using both Stickman Pose Estimation and the amendment by PGO. We conducted experiments with 10-fold cross validation. Then, we also explain scenario (c) for evaluation of the energy function in Sect. 6.4, and scenario (d) and (e) for PGO refinement in Sect. 7.

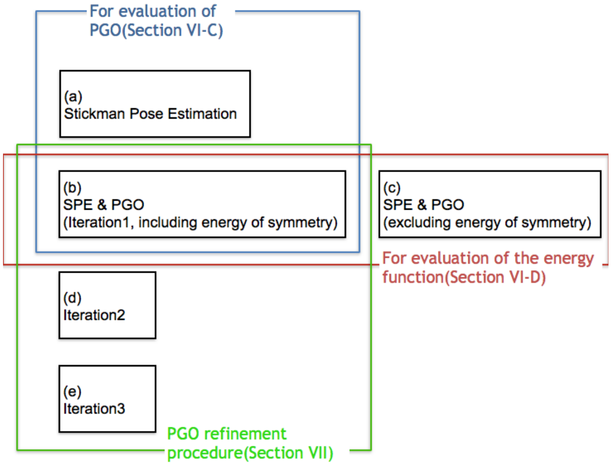

For evaluation of PGO(Section VI-C)
(a) Stickman Pose Estimation
(b) SPE & PGO (Iteration1, including energy of symmetry)
(c) SPE & PGO (excluding energy of symmetry)
For evaluation of the energy function(Section VI-D)
(d) Iteration2
(e) Iteration3
PGO refinement procedure(Section VII)

Fig. 8 Five different scenarios

6.1 Dataset

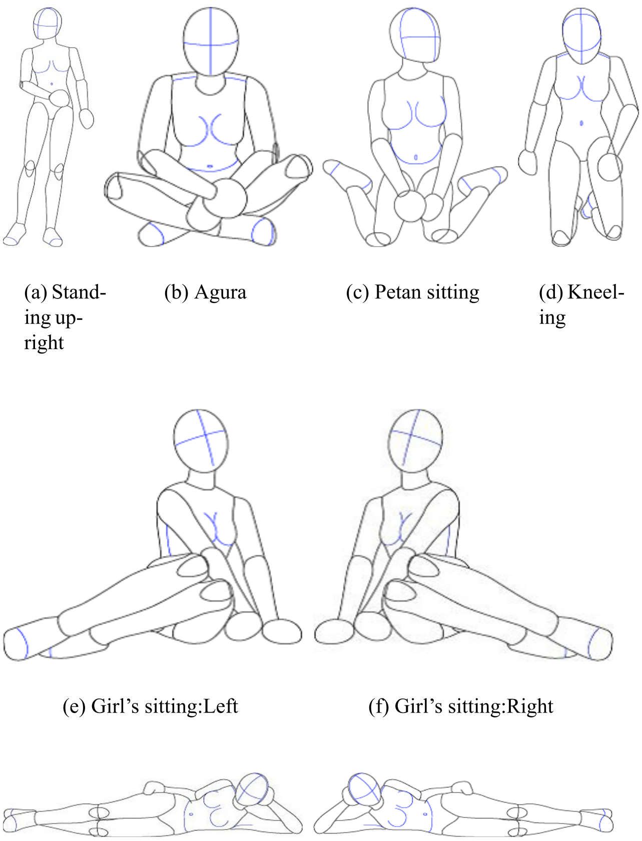
First, we introduce a dataset used to train and test the classifiers in our method. We chose the following 8 poses which are frequently found in Japanese idol photos, since more than 70% images obtained by Google image search with queries of idol names are divided into the following 8 poses, except for the images of the upper body only. Description and criteria of each pose is explained below. We selected images which fit to the criteria as dataset for each category, and three of the members in our laboratory confirmed if there is no deviation in categories. Also, we made the example figures for each category open to the public on our website[††].

1. Standing Upright (Fig. 9 (a))
   A person stands facing almost the front, or obliquely, except for completely turning back. There is no limitation on the arms and legs positions, including crossing and/or bending the elbows and the knees.
2. Agura (Fig. 9 (b))
   A person sits crossing the legs in the front, facing the front or obliquely, except for completely turning back. There is no limitation on the arms positions.
3. Petan sitting (Fig. 9 (c))
   A person sits bending the legs on either side, facing the front or obliquely, except for completely turning back. There is no limitation on the arms positions. The angles of the knees should be under 90 degrees.
4. Kneeling (Fig. 9 (d))
   A person puts the knees on the ground without bending the waist, facing the front or obliquely, except for completely turning back. There is no limitation on the arms positions and the direction of the lower legs.
5. Girl's sitting (Fig. 9 (e), (f))
   A person sits bending the legs on the right or left, facing the front or obliquely, except for completely turning back. There is no limitation on the arms positions. The angles of the knees should be under 90 degrees.
6. Lying (Fig. 9 (g), (h))
   A person puts the waist and the right or left knee on the ground, making almost all the parts visible. There is no limitation on the arms positions and the knees's angles.

The dataset contains 400 photos. Each class has 50

[†] http://www.cs.waikato.ac.nz/ml/weka

[††] www.ohsuga.is.uec.ac.jp/~kawamura/poses.zip



(a) Standing upright (b) Agura (c) Petan sitting (d) Kneeling

(e) Girl's sitting:Left (f) Girl's sitting:Right

(g) Lying:Left (h) Lying:Right

Fig. 9 Dataset

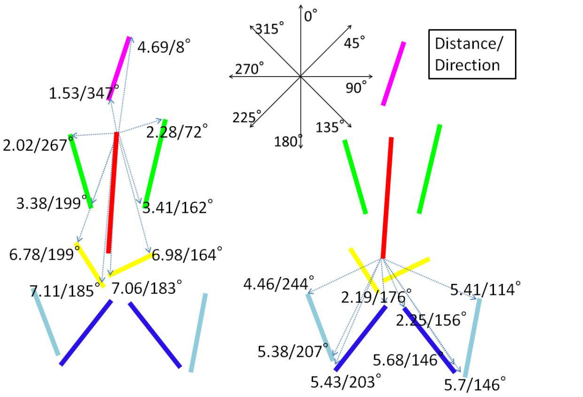


Fig. 10 Feature 1 and 2: Distance and direction between Torso and other parts. Feature 1 and 2 have 18 values for each, because 9 body parts, except for torso, has a start point and an end point.

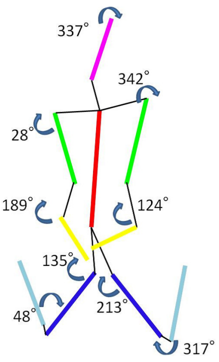


Fig. 11 Feature 3: Angle between two body parts that have the connection. Feature 3 has 9 values, because human body has 9 joints between body parts.

images and all images contain a single person. We should note that our method confirms whether there is a person in an image by detecting a face. Thus, in the case of the face detection failure, the method does not go to the next procedure. However, in the case of a face in profile, if a face is detected, the method can classify his/her poses. Images which have no face (or head) are out of scope of our research.

6.2 Feature Selection

In this paper, we assumed that a pose is determined by the relative position of each body part (Fig. 10) and angles between two parts (Fig. 11). Therefore, we selected three kinds of features as input for the Bayesian Network classifier. In the first experiment, we used a feature vector which has 45 dimensions of numerial features. Feature values which represent the distances of parts are normalized by dividing by the length of the torso, and feature values which represent the angles are not normalized, since they have upper and lower bounds.

- Feature 1: Distance between torso and other parts (Fig. 10), which has 18 dimensions.
- Feature 2: Direction from torso to other parts (Fig. 10), which has 18 dimensions.
- Feature 3: Angle between two body parts that have a connection (Fig. 11), which has 9 dimensions.

6.3 Result and Analysis

Table 1 (a) and (b) shows precision, recall and F-measure of each pose in two scenarios. As shown in the table, the F-measure of the classification is about 9% higher than the non-amendment. Amended by the PGO, the accuracy of almost all the poses has been improved. Although the amendment by the PGO improves the accuracy of the classification, the F-measure of Agura, Petan sitting, and Kneeling are relatively low, about 60%, and there is room for further improvement. Therefore, we propose an iterative ontology refinement procedure. Table 2 shows the results of classification on Iteration 1 (the experiment that use both Stickman Pose Estimation and the amendment by PGO).

Also, these poses have some common shapes of the body parts, which make harder to distinguish them. Especially, 13 out of 50 images of Petan sitting are incorrectly labeled as Agura.

6.4 Experiment on Evaluation of the Energy Function

We also conducted experiments for evaluation of the energy

Table 1 Precision, Recall, and F-measure of each pose

(a) SPE (Iteration 1)

	Pose	Precision	Recall	F-measure
SPE Iteration 1	Standing upright	96.1%	98.0%	97.0%
	Agura	47.1%	48.0%	47.5%
	Petan sitting	44.0%	44.0%	44.0%
	Kneeling	58.5%	62.0%	60.2%
	Girl's sitting:Left	77.3%	68.0%	72.3%
	Girl's sitting:Right	68.6%	70.0%	69.3%
	Lying:Left	97.9%	94.0%	95.9%
	Lying:Right	94.2%	98.0%	96.1%
	Average	**73.0%**	**72.8%**	**72.8%**

(b) SPE&PGO (Iteration 1, including energy of symmetry)

	Pose	Precision	Recall	F-measure
SPE & PGO Iteration 1	Standing upright	92.5%	98.0%	95.1%
	Agura	60.0%	66.0%	62.9%
	Petan sitting	76.9%	60.0%	67.4%
	Kneeling	59.3%	64.0%	61.5%
	Girl's sitting:Left	87.5%	84.0%	85.7%
	Girl's sitting:Right	85.4%	82.0%	83.7%
	Lying:Left	98.0%	98.0%	98.0%
	Lying:Right	94.3%	100.0%	97.1%
	Average	**81.7%**	**81.5%**	**81.4%**

(c) SPE&PGO (excluding energy of symmetry)

	Pose	Precision	Recall	F-measure
SPE & PGO excluding symmetry	Standing upright	92.6%	100.0%	96.2%
	Agura	45.7%	42.0%	43.7%
	Petan sitting	55.0%	44.0%	48.9%
	Kneeling	60.7%	74.0%	66.7%
	Girl's sitting:Left	60.8%	62.0%	61.4%
	Girl's sitting:Right	61.7%	58.0%	59.8%
	Lying:Left	100.0%	96.0%	98.0%
	Lying:Right	94.3%	100.0%	97.1%
	Average	**71.3%**	**72.0%**	**71.5%**

(d) Iteration 2

	Pose	Precision	Recall	F-measure
Iteration 2	Standing upright	98.0%	98.0%	98.0%
	Agura	80.4%	74.0%	77.1%
	Petan sitting	70.6%	72.0%	71.3%
	Kneeling	63.3%	62.0%	62.6%
	Girl's sitting:Left	91.8%	90.0%	90.9%
	Girl's sitting:Right	83.3%	90.0%	86.5%
	Lying:Left	98.0%	98.0%	98.0%
	Lying:Right	98.0%	100.0%	99.0%
	Average	**85.4%**	**85.5%**	**85.4%**

(e) Iteration 3

	Pose	Precision	Recall	F-measure
Iteration 3	Standing upright	100.0%	100.0%	100.0%
	Agura	83.3%	80.0%	81.6%
	Petan sitting	76.8%	86.0%	81.1%
	Kneeling	69.0%	58.0%	63.0%
	Girl's sitting:Left	88.7%	94.0%	91.3%
	Girl's sitting:Right	84.3%	86.0%	85.1%
	Lying:Left	100.0%	98.0%	99.0%
	Lying:Right	98.0%	100.0%	99.0%
	Average	**87.5%**	**87.8%**	**87.5%**

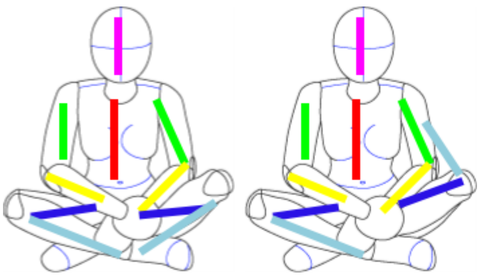
Table 2 Iteration 1: Result of classification. The first column contains the 8 poses for classification and the last eight columns show how images of dataset are classified. The numbers of images correctly classified are described in bold type.

	a	b	c	d	e	f	g	h
a.Standing upright	**49**	0	0	1	0	0	0	0
b.Agura	0	**33**	1	9	3	3	0	1
c.Petan sitting	1	13	**30**	5	1	0	0	0
d.Kneeling	2	8	3	**32**	1	4	0	0
e.Girl's sitting:Left	1	1	3	2	**42**	0	1	0
f.Girl's sitting:Right	0	0	2	5	0	**41**	0	2
g.Lying:Left	0	0	0	0	1	0	**49**	0
h.Lying:Right	0	0	0	0	0	0	0	**50**

function detailed in Sect. 5.2. In the above section, we assumed the experiment using the energy function E in Eq. (1), but in this section, we use the energy function, excluding symmetry energy in Eq. (3) for comparison. This makes difference in amending procedures between two scenarios. The former procedure is showed in Sect. 5.2.4, and the latter is showed as follows.

1. Evaluate the result of Stickman Pose Estimation, and detect constraints violation.
2. If any limbs violated the constraints provided by PGO, amendment procedure is executed (Upper arms or legs, which do not violate, are amended only if lower arms or legs violated).
3. 5x5 grid points, which are candidates for start points of the upper arms or legs, are arranged in the designated area. The width and height of the area are set as the same as length of torso. The area is placed at just beside a start point of torso when arms are amended. When legs are amended, it is placed at diagonal downward from an end point of torso.
4. Upper parts are located around each grid point at 10 degrees interval, and the color information of surrounding pixels is taken at each location. For reducing calculation cost, lower 50% of parts location with relatively small skin-color pixels are removed.
5. Also, lower parts are located around each end point of upper parts, and the color information of surrounding pixels is taken at each location.
6. The candidate of part including most skin-color pixels is selected.
7. If the selected parts violate the constraints provided by PGO, the second most skin-colored one is adopted.

Table 1 (c) shows precision, recall and F-measure of each pose on this experiment. As shown in the table, the F-measure of the classification is not only 10% lower than SPE & PGO including energy of symmetry, but also lower than just SPE. Amendment excluding the energy of symmetry caused inappropriate computation of position, since it depends only on skin-color information in a large area for start points of parts. As Fig. 12 (b) shows, a leg is mistakenly amended as an arm, and thus the result becomes Girl's sitting in many cases, although the pose is Agura. This typical failure occurs frequently, leading to decrease the accuracy of Girl's sitting. This result shows that the energy of symmetry correctly amends the start points of upper parts and effects the accuracy of classification.



(a) Including energy of symmetry

(b) Excluding

Fig. 12 Estimation results of experiments, including energy of symmetry and excluding it.

7. PGO Refinement Procedure

In the above section, we assumed that the human pose is determined by the relative position of each part and angle between two parts, and thus PGO generates the above-mentioned three kinds of features. In the following sections, we repeatedly refine the PGO by confirming the classification results.

7.1 Feature Selection for Iteration 2

From Iteration 1 results and analysis, we derived the Feature 4 (Fig. 13), following additional features, that could contribute to discriminate the three poses of low accuracy.

- Feature 4: Absolute angle of body parts (Fig. 13), which has 10 dimensions.

The absolute angle of lower legs obviously distinguishes Agura and Petan sitting (Fig. 13), and will contribute to improve the accuracy of classification. The additional features, absolute angle of body parts are described in red as "Cla:AbsAngle" in Fig. 5.

7.2 Experiment for Iteration 2

Table 1 (d) shows precision, recall and F-measure of each pose on Iteration 2. As shown in the table, the F-measure of the classification is 4% higher than Iteration 1.

Table 3 shows the results of the classification on Iteration 2. The number of images of Petan sitting which are incorrectly labeled as Agura, decreased from 13 in Iteration 1 to 3 in Iteration 2. This result suggests that Feature 4 has been effective in the pose classification.

7.3 Feature Selection for Iteration 3

Considering the result of Iteration 2, we further refine PGO for the next experiment (Iteration 3) to improve the accuracy of the classification. In Iteration 3, we derived Feature 5 and 6 (described in Fig. 14), that represent the relative positions of upper legs and lower legs in the same way as Iteration 2.

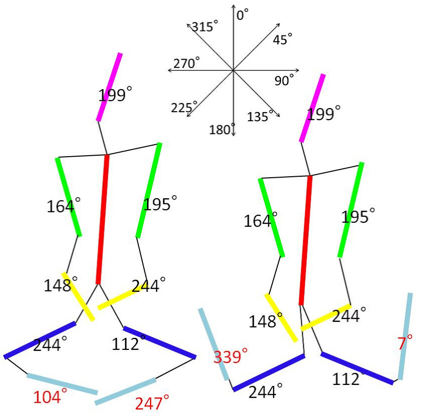


Fig. 13 Feature 4: Absolute angle of body parts, obviously distinguishes Agura and Petan sitting

Table 3 Iteration 2: Result of classification

	a	b	c	d	e	f	g	h
a.Standing upright	**49**	0	0	1	0	0	0	0
b.Agura	0	**37**	4	7	0	2	0	0
c.Petan sitting	0	3	**36**	8	1	2	0	0
d.Kneeling	1	5	7	**31**	2	4	0	0
e.Girl's sitting:Left	0	1	2	0	**45**	1	1	0
f.Girl's sitting:Right	0	0	2	2	0	**45**	0	1
g.Lying:Left	0	0	0	0	1	0	**49**	0
h.Lying:Right	0	0	0	0	0	0	0	**50**

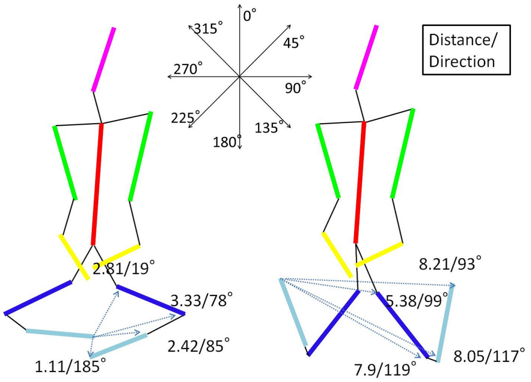


Fig. 14 Feature 5 and 6: distance and direction between the start and the end points of the left and right upper and lower legs. They have 16 values respectively, because left and right legs have 4 points. Feature 5, 6 obviously distinguishes Agura and Petan sitting

- Feature 5: Distance between the start and the end points of the left and right upper and lower legs (Fig. 14), which has 16 dimensions.
- Feature 6: Direction between the start and the end points of the left and right upper and lower legs (Fig. 14), which has 16 dimensions.

To simplify the figure, Feature 5 and 6 are described as the relations between the end point of left lower leg and both the end points of right legs. However, in the actual calculation, they mean relations between both the end points of left leg and the ones of right leg.

Figure 14 expresses that distance and direction between

Table 4 Iteration 3: Result of classification

	a	b	c	d	e	f	g	h
a.Standing upright	**50**	0	0	0	0	0	0	0
b.Agura	0	**40**	3	6	0	1	0	0
c.Petan sitting	0	1	**43**	5	1	0	0	0
d.Kneeling	0	6	4	**29**	4	7	0	0
e.Girl's sitting:Left	0	1	1	1	**47**	0	0	0
f.Girl's sitting:Right	0	0	5	1	0	**43**	0	1
g.Lying:Left	0	0	0	0	1	0	**49**	0
h.Lying:Right	0	0	0	0	0	0	0	**50**

the end point of left lower leg and both the end points of right leg obviously distinguishes Agura and Petan sitting. The new features, such as distance and direction between the start points and the end points of the left and right upper and lower legs, are described in green as "Cla:DisOfEnd2End" and "Cla:DirOfEnd2End" at Leg Relation class in Fig. 5.

7.4 Experiment for Iteration 3

Table 1 (e) shows precision, recall, and F-measure of each pose on Iteration 3. As shown in Table 1 (e), the F-measure of the classification is 2.25% higher than Iteration 2.

7.5 Evaluation

In Iteration 3, the classification accuracy of Agura and Petan sitting is improved as expected. On the other hand, the recall of Kneeling decreased during the above iterative processes. A reason for this result is that the classifier cannot capture the features of Kneeling and learn them. Kneeling may have left, right, or front direction, so that the feature of Kneeling has become a mixture of these three directions.

Especially, Girl's sitting: left and right have features similar to those of Kneeling: left and right. While the classifier cannot capture the feature of Kneeling, that of Girl's sitting is learned correctly. Consequently, the number of images of Kneeling that are incorrectly labeled as Girl's sitting, increased in each Iteration (Table 4).

8. Conclusion and Future Work

Although we estimated and classified human poses from a 2d- still image, most of the recent researches on pose classification handle moves and more depth images, and then there is little research for the a single still image. However, since 2d- still images are most popular formats in the Internet, we consider that handling of the still images has the significance in the CV area. Then, the research we used as reference is Eichner's work [7], which achieved around 80% accuracy depending on background settings for estimation of body parts, assuming that the target images have a front or rear upper half of the body. On the other hand, we achieved more than 80% accuracy of pose classification. We have no limitation on the background settings, and the target images can have the whole body from several angles, which are collected from the Internet. Although we assume that an image has a single person, by splitting an image per person in a pre-process, we can extend our method to an image with multiple persons. In terms of the problem of various textures, we are considering to solve it with a method proposed by the related works to [7]. Thus, although we could not directly compare the referenced work due to different conditions, we consider that our contribution with different (ontological) approach is similar to the top research in a meaningful research topic. Detailed contributions of this paper include:

1. We improved the stickman pose estimation method, in order to deal with Japanese idol poses.
2. We proposed and exploited Pose Guide Ontology and energy functions to amend the result of stickman pose estimation. PGO contains constraints and semantic hierarchies related to the orientation and positional relationship between the body parts. The spatial layout of the body parts are optimized so that energy functions are minimized.
3. We also proposed an iterative procedure for further refinements of PGO to effectively classify the poses. Considering the results of experiments, we added properties to PGO, which are features that distinguish the poses. The final results indicated that F-measure of the classification has become 15% higher than non-amended results.

In order to further improve the proposed method, we plan to subdivide the pose types and define the new constraints. As described in Sect. 7.5, the dataset in the experiment did not divide Kneeling into three poses (right, left, and front). As a result, the features of Kneeling has become a mixture of these three directions. Therefore, by further division of the pose types, the classifier could capture the features and learn the poses correctly.

Also, the constraints which are currently defined by PGO do not amend all the (incorrect) parts information obtained by the pose estimation. However, too strict constraints tend to amend the right result of estimation incorrectly. So that, we will define the effective constraints and integrate them into the existing ones instead of single strict constraints.

Currently, our method takes about 40 (sec) to process an image. The breakdown of that is image loading and face detection: 25%, foreground highlighting: 35% and parts estimation: 40%. When the result of parts estimation requires correction of PGO, it further takes 20–40 (sec). We intend to address the reduction of the processing time as a future work, although the classification should be offline.

In the near future, we would like to challenge more complex, and socially significant problems like image analysis of security cameras and for rehabilitation exercises of patients. We plan to further extend this research, and contribute the situations, to which the conventional methods cannot be applied.

References

[1] A. Gepperth, M.G. Ortiz, and B. Heisele, “Real-time pedestrian detection and pose classification on a GPU,” 16th International IEEE Conference on Intelligent Transportation Systems, pp.348–353, 2013.

[2] S. Mukherjee, S.K. Biswas, and D.P. Mukherjee, “Recognizing interaction between human performers using ’key pose doublet’,” Proceedings of the 19th ACM International Conference on Multimedia, MM ’11, New York, NY, USA, pp.1329–1332, ACM, 2011.

[3] G. Mori, X. Ren, A.A. Efros, and J. Malik, “Recovering human body configurations: Combining segmentation and recognition,” Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR’04, Washington, DC, USA, pp.326–333, IEEE Computer Society, 2004.

[4] X. Zhang, M. Ye, Y. Xu, and Z. Tian, “Locally adaptive combining colour and depth for human body contour tracking using level set method,” IET Computer Vision, IET, vol.8, no.4, pp.316–328, 2014.

[5] M.A. Fischler and R.A. Elschlager, “The representation and matching of pictorial structures,” IEEE Trans. Comput., vol.22, no.1, pp.67–92, Jan. 1973.

[6] D. Ramanan, “Learning to parse images of articulated bodies,” NIPS’06, pp.1129–1136, 2006.

[7] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, “2d articulated human pose estimation and retrieval in (almost) unconstrained still images,” Int. J. Comput. Vision, vol.99, no.2, pp.190–214, Sept. 2012.

[8] M. Marszalek and C. Schmid, “Semantic hierarchies for visual object recognition,” CVPR, IEEE Computer Society, pp.1–7, 2007.

[9] G.A. Miller, “Wordnet: A lexical database for english,” Commun. ACM, vol.38, no.11, pp.39–41, Nov. 1995.

[10] I. Nwogu, V. Govindaraju, and C. Brown, “Syntactic image parsing using ontology and semantic descriptions,” CVPR, IEEE Computer Society, pp.41–48, 2010.

[11] N. Chen, Q.-Y. Zhou, and V. Prasanna, “Understanding web images by object relation network,” Proceedings of the 21st international conference on World Wide Web, WWW ’12, New York, NY, USA, pp.291–300, ACM, 2012.

[12] C. Rother, V. Kolmogorov, and A. Blake, “”grabcut”: interactive foreground extraction using iterated graph cuts,” ACM Trans. Graph., vol.23, no.3, pp.309–314, Aug. 2004.

[13] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, “Progressive search space reduction for human pose estimation,” Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp.1–8, 2008.

[14] P.F. Felzenszwalb and D.P. Huttenlocher, “Pictorial structures for object recognition,” IJCV, vol.61, vol.61, no.1, pp.55–79, 2003.

Kazuhiro Tashiro received his M.S. degree in computer science from the University of Electro-Communications in 2014. He joined CyberAgent, Inc in 2014. His research interests include the Semantic Web, and ontology.

Takahiro Kawamura received his Ph.D. degree in computer science from Waseda University in 2001. He was a visiting researcher in Carnegie Mellon University, US, from 2001 to 2002. He is a senior researcher in Japan Science and Technology Agency. He is also an visiting associate professor in the Graduate School of Information Systems at the University of Electro-Communications, and a part-time lecturer in the Graduate School of Engineering at Osaka University. His research interests include the Semantic Web, open data, and software agent. He was a board member of the Japanese Society for Artificial Intelligence (JSAI) from 2012 to 2013.

Yuichi Sei received his Ph.D. degree in Information Science and Technology, from the University of Tokyo, Japan, in 2009. From 2009 to 2012 he was working at Mitsubishi Research Institute. He is an assistant professor at the University of Electro-Communications. His research interests include pervasive computing, security, and privacy-preserving data mining.

Hiroyuki Nakagawa recived his Ph.D. degree in computer science from Waseda University in 2013. He worked in Kajima Corporation from 1997 to 2008 and at the University of Electro-Communications as an assistant professor from 2008 to 2013. He is an associate professor in the Graduate School of Information Science and Technology, Osaka University. His research interests include requirements engineering, self-adaptive systems, and agent technologies.

Yasuyuki Tahara received his PhD in Information and Computer Science from Waseda University, Japan, in 2003. He was a visiting researcher in City University London, UK, from 1995 to 1996, and in Imperial College London, UK, from 1996 to 1997. He is an associate professor in the University of Electro-Communications. His research interests include formal verification of software and requirements engineering.

Akihiko Ohsuga received his a Ph.D. degree in computer science from Waseda University in 1995. He is a professor in the Graduate School of Information Systems, the University of Electro-Communications (UEC). He is also a visiting professor in National Institute of Informatics (NII). His research interests include agent technologies, web intelligence, and software engineering. He is a member of the IEEE Computer Society, the Information Processing Society of Japan (IPSJ), the Institute of Electronics, Information and Communication Engineers (IEICE), The Japanese Society for Artificial Intelligence (JSAI), and Japan Society for Software Science and Technology (JSSST).