A hierarchical extension of the HOG model
implemented in the convolution-net for human detection

Yasuto Arakaki ,[†1] Hayaru Shouno ,[†1] Kazuyuki Takahashi [†2]
and Takashi Morie [†2]

For the detection of generic objects in the field of image processing, histograms of orientation gradients (HOG) is discussed for these years. The performance of the classification system using HOG shows a good result. However, the performance of using HOG descriptor would be influenced by the detecting object size. In order to overcome this problem, we introduce a kind of hierarchy inspired from the convolution-net, which is a model of our visual processing system in the brain. The hierarchical HOG (H-HOG) integrates several scales of HOG descriptors in its architecture, and represents the input image as the combinatorial of more complex features rather than that of the orientation gradients. We investigate the H-HOG performance and compare with the conventional HOG. In the result, we obtain the better performance rather than the conventional HOG. Especially the size of representation dimension is much smaller than the conventional HOG without reducing the detecting performance.

1. Introduction

In the field of generic object recognition, the histogram of oriented gradient (HOG) proposed by Dalal has been focused for description of objects in the images because of its simple feature extraction rule[1,2]. HOG represents features of an object as a histogram of the input image gradient of certain areas that includes the object. Using HOG for the recognition of the image area, the histogram of the image is usually treated as a vector for the input of classification machine, such like a support vector machine (SVM). Even though HOG is a simple model, it has shown several good recognition performances for pedestrian and car detection[3].

However, we consider HOG includes two problems to solve. The first point is the number of feature dimensions, which can become huge number by the value of image dividing parameter. The large number of feature dimensions prevents reducing computational cost and requires a lot of training images for classification, so that we should

---

†1 University of Electro-Communications
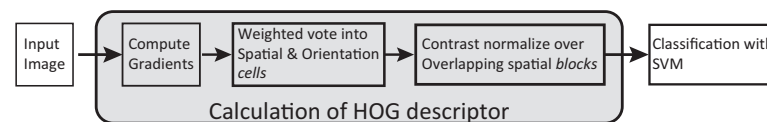†2 Kyushu Institute of Technology

Fig. 1 A schematic diagram of a typical image recognition system using HOG descriptor[1].

reduce it for the cost down. The second point is that invariance for both the location and the scale of the object may not be good enough for the robust recognition.

On the other hand, we human have a flexible recognition system, for example, we can recognize an object in the image with any locations and with several scale changes. From the physiological viewpoints, our visual processing system in the brain is believed to have a kind of hierarchical structure[4,5]. HOG can be interpreted as a simple layered structure neural network, so that we consider introducing hierarchical structure such like the brain into the HOG can improve the robustness against the deformation of the objects. These kinds of hierarchical neural networks are called "convolution-net"[6,7,8]. In the convolution-net, one of the important points is hierarchy, so that we introduce a kind of hierarchical structure represented by the convolution-net into the HOG. Introducing the hierarchy, we expect the robustness for variations of both location and scale can be improved to the conventional HOG model. In this study, we proposed an improving model of the HOG that introduce a concept of the hierarchical structure of the convolution-net. In this study, we evaluate the performance of our proposing model by use of the INRIA Person Dataset, which is an image database for the pedestrian detection, and we discussed about recognition performance.

2. Conventional Model Formulation

In this study, we introduce the hierarchical structure, which is inspired from visual processing in our brain, into the conventional HOG called Hierarchical HOG (H-HOG). Thus, we explain summaries about conventional visual recognition system using HOG, and the convolution-net proposed by Mutch & Lowe in this section[1,9].

2.1 Conventional HOG system

The conventional HOG, which is applied to the generic image recognition, is a kind of feature descriptor using histograms of the local image gradients. We can obtain a HOG descriptor, which can represent a rough shape of the object, into the local area of an image, so that the HOG descriptor is often applied to the object detection such like

human detection[3].

Fig.1 shows the schematic diagram of a typical visual recognition system using the HOG descriptors. At first, the image gradient is calculated from the input image $I(u, v)$, where (u, v) indicates the location in the image and $I(u, v)$ means the pixel value at the location. The gradient for each location (u, v) is represented as both intensity $m(u, v)$ and orientation $\theta(u, v)$,

$$m(u, v) = \sqrt{I_u(u, v)^2 + I_v(u, v)^2}, \tag{1}$$

$$\theta(u, v) = \tan^{-1}\left(\frac{I_v(u, v)}{I_u(u, v)}\right), \tag{2}$$

where $I_u(u, v)$ and $I_v(u, v)$ means the difference between the neighbors:

$$I_u(u, v) = I(u + 1, v) - I(u - 1, v) \tag{3}$$

$$I_v(u, v) = I(u, v + 1) - I(u, v - 1). \tag{4}$$

2.1.1 Histogram representation for the cell

In the next step for obtaining HOG descriptor, we divide the gradient into several small areas called "cell" and make an orientation histogram for each cell. In the process, the intensity $m(u, v)$ plays a roll of the weight for the voting bin of $\theta(u, v)$. Applying such local orientation histograms, local translation deformation effect of the object in the cell is reduced. The histogram, which describes the distribution of the edge components for the orientation, can be regarded as a vector. Quantizing the orientation $\theta(u, v)$ on the ith cell into the Q state, we can obtain Q bins histogram whose elements are represented by the vector: $\boldsymbol{f}_i = \{f_{i,1}, f_{i,2}, \cdots, f_{i,Q}\}$:

$$f_{i,q} = \sum_{(u,v)\in i\text{th cell}} m(u, v)\, \delta_q(\theta(u, v)), \tag{5}$$

$$\delta_q(s) = \begin{cases} 1 & (q - 1)\Delta s \leq s < q\Delta s \text{ where } \Delta s = \pi/Q \\ 0 & \text{else} \end{cases}. \tag{6}$$

The size of cells is a important parameter that makes influence to the number of feature dimension for the classifier. Large cell size conducts small number of the feature dimension, which is treatable property for the classifier; however, feature extraction may become too rough to represent the object. On the contrary, small size cell size, which can represent detail of the object, may make huge number of feature dimensions to describe the object, which makes hard problem for the classifier.

2.1.2 Block Normalization for the cells representation

The gradient strength varies over a wide range owing to local variations in illumination and contrast between foreground and background. Thus, effective local contrast normalization may be good for the classification. We adopt L2 normalization in the same manner of Dalal[1]. Considering the neighboring cells for the ith cell as a group, which is called "block", the ith block feature can be described as collection of the histogram vectors:

$$\tilde{\boldsymbol{V}}_i = \{\boldsymbol{f}_i, \{\boldsymbol{f}_k\}_{k\in\text{NN}(i)}\}, \tag{7}$$

where NN(i) means the neighbor cells for the ith cell. Then the block vector $\tilde{\boldsymbol{V}}_i$ is normalized as

$$\boldsymbol{V}_i = \frac{\tilde{\boldsymbol{V}}_i}{\sqrt{\|\tilde{\boldsymbol{V}}_i\|^2 + \epsilon^2}}, \tag{8}$$

where ϵ is a small positive constant to prevent diverging. HOG descriptor, in the final form, is a collection of these normalized block vectors $\{\boldsymbol{V}_i\}$.

2.1.3 Conventional HOG classifying system

HOG descriptor is a robust expression for the local translation deformation and illumination variation, and it can represent rough feature of the object in an image. Thus, HOG is considered as good for the generic object recognition[2]. Dalal proposed to use HOG descriptor as the input for the support vector machine (SVM), which is a kind of classifier, and showed better classification performance rather than those of other features[1].

2.2 Summary of a Convolution-net of Mutch & Lowe

For introducing the brain inspired mechanism into the HOG, we explain a kind of convolution-net proposed by Mutch & Lowe[9]. The convolution-net is a hierarchical artificial neural network model originated from the model proposed by Hubel and Wiesel[5]. Hubel & Wiesel found two types of cells in the early visual processing area of the mammal brain, which are called "simple cell" and "complex cell". Each type of cell has local area for responding in the viewing field, which is called "receptive field", and responds to the specific input stimulus such that line/edge component in the receptive field. The difference between these types of cells is response for the location of the preferred input stimulus in the receptive field. The simple cell only responds to the preferred input stimulus at the specific location. On the other hand, the complex cell responds to the preferred input stimulus at any location in the receptive field. Thus, Hubel & Wiesel proposed a kind of hierarchy between these cells, that is, a complex cell may gather the outputs of simple cells that respond to same preferred stimulus but have slightly different receptive fields. The convolution-net has a hierarchy of these types of cells, and connects this hierarchy alternately[6][7][9].

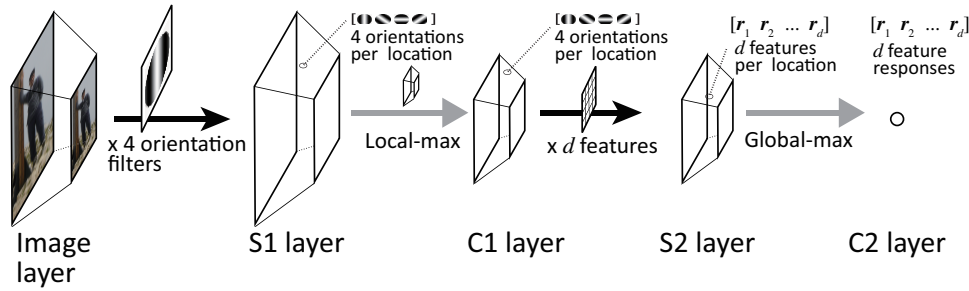Fig.2 is a schematic diagram of the convolution-net proposed by Mutch & Lowe[9].

Fig. 2 A schematic diagram of the convolution-net proposed by Mutch&Lowe[9]. The input image would be processed through the pathway, that is, S1 → C1 → S2 → C2 layers. The input image is translated the description of d dimensional vector appeared in $C2$ layer.
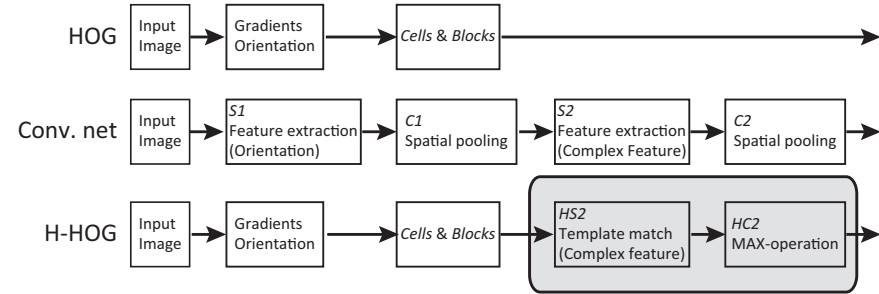

Fig. 3 Comparison among conventional HOG, convolution-net, and our hierarchical HOG. The convolution-net alternate S-layer, which extracts features, and C-layer, which tolerate deformation with spatial pooling. The conventional HOG corresponds to the S1-layer and C1-layer in the convolution-net. Our hierarchical HOG extends HOG description with the manner of convolution-net, which is described as S2 and C2 layers.

The lowest of the figure shows the input layer. The most specific feature of this model is introducing the multi-scale expression for the input image. An input image is scaled into the several resolutions, which is described as the image pyramid in the bottom of the figure. The S1-layer in the figure is a model for the simple cells that extract line/edge components for each location and resolution. Mathematically, this extracting operation can be described as a convolution with line/edge filtering template. The next C1-layer in the figure is a model for the complex cells. The C1-layer cell calculate local maximum for the corresponding S1-layer cells. This type of operation is called spatial pooling, which is to tolerate the deformation of local translation of the pattern in the image.

The S2-layer is a higher feature extraction layer, which is implemented as a kind of template-matching mechanism. The S2-layer consists of d types of templates, and each template is also treated as the line/edge extracting filters. These template filters are obtained with sampling from the patterns appeared in C1-layer for training input images.

The final layer, which is called C2-layer, integrates extracted features in the S2-layer pyramids. The unit in the C2-layer detects maximum value for the corresponding S2-layer pyramid. As the result, the unit represents the containing rate of the template pattern for the input pattern. The expression of the C2 layer can be regarded as the d dimensional vector for an input image, so that, we can apply several classifiers such like SVM for this expression.

3. Hierarchical HOG formulation

The hierarchical architecture of the convolution-net is an important concept for our

study. The S1 layer in the convolution-net plays a role of line/edge extractor for in the viewing field. Calculating histogram from the image gradient operation in HOG is the similar function for the line/edge extractor. Focusing to the function of line/edge extraction, the difference between these two models is only representation for the extracted features. The function of the C1 layer and that of the cell/block mechanism in HOG is also similar. Fig.3 shows the corresponding architectures between the convolution-net and HOG. In the figure, the conventional HOG output corresponds to the C1 layer output; however, the convolution-net has more deep hierarchy such like S2 and C2 layer. Thus, we can introduce hierarchy extension for HOG in the manner of the convolution-net. In the following we call our hierarchical HOG as "H-HOG".

3.1 Input feature modulation

Before explaining the higher layers, we introduce several modifications for the calculation of HOG descriptor. In the fig.3 HOG descriptor, which is conducted from eq.(6) to eq.(8), is calculated as the input of higher stage. In the count up for the histogram of eq.(6), the intensity $m(u, v)$ is piled up linearly, however, the small value of $m(u, v)$ may be a kind of contamination. Considering the case that all components of ith block $\tilde{V}_i$ are weak, all the weak components are enhanced in the normalization procedure in eq.(8). Thus, we introduce a nonlinear modulation in order to emphasize the histogram:

$$f_{i,q} = \sum_{(u,v) \in i\text{th cell}} \psi(m(u, v))\, \delta_q(\theta(u, v)), \tag{9}$$

$$\tag{10}$$

$$\psi(x) = \begin{cases} |x - h|^d & \text{if } x > h \\ 0 & \text{else} \end{cases}, \tag{11}$$

where h plays a role of threshold and d means a enhance factor. After calculating the modulated histograms $\{f_i\}$, we adopt the same manner of the conventional HOG representation.

3.2 Multi-resolution representation

The convolution-net proposed by Mutch & Lowe introduces a multi-resolution representation. In order to introduce the multi-resolution representation in our system, we prepare several HOG descriptors that have different cell sizes.

3.3 template-matching layer: $HS2$

The template-matching layer in the H-HOG corresponds to the S2 layer in the convolution-net, which is to extract just more complex feature rather than those of the S1 layers. Mutch&Lowe adopt to use templates of partial C1 descriptions for the training patterns chosen by sampling. For convenience, we call this template-matching layer as $HS2$-layer.

3.3.1 Template selection

In the conventional HOG, the output description is represented by $\{V_i\}$ in eq.(8), where i means the block location. In the training mode of the H-HOG, at first, we select several block locations in the several resolutions randomly. We treat neighbor blocks of the selected block as a cluster for the template, and the size of cluster is selected from several variations randomly. Thus, when location i' is selected for the kth template T_k, the template vector can be denoted as:

$$T_k = \{V_{i'}, \{V_j\}_{j \in \mathrm{NN}_k(i')}\}, \tag{12}$$

where $\mathrm{NN}_k(i')$ means collection of neighborhood blocks of the i'th block, and the size of the neighborhood is chosen from several variations randomly.

Moreover, in order to prevent using similar template, we adopt the following discard rule. Selecting new template T_n from the input description, we calculate similarities for whole existing templates by use of direction cosine:

$$u_{n,k} = \frac{T_n \cdot T_k}{\|T_n\| \, \|T_k\|}, \tag{13}$$

where k is the index for the any existing templates. If the similarity $u_{n,k}$ is larger than a threshold U, we regard the $HS2$ layer already has template T_n and the template T_n is not accepted for the $HS2$ layer.

3.3.2 Calculation of $HS2$ layer representation

For the representation of the input image in the $HS2$ layer, we adopt direction cosine between a template and input representation in for the feature extraction of the $HS2$ layer. Denoting X_j as the conventional HOG description of the location j for the kth template, that is,

$$X_j = \{V_j, \{V_l\}_{l \in N_k(j)}\}, \tag{14}$$

we describe the output of the $HS2$ layer $R_{k,j}^{HS2}$ as

$$R_{k,j}^{HS2} = \frac{T_k \cdot X_j}{\|T_k\| \, \|X_j\|}. \tag{15}$$

Thus, the feature description $R_{k,j}^{HS2}$ describes the including degree of kth template at the location j. Applying this operation to whole input location, we can obtain a feature map for the kth template.

3.4 Max-operator layer: HC2

In the convolution-net, the C2 layer plays a role of integration of feature by use of maximum operation, which can be considered as a kind of spatial pooling. In the H-HOG model, we also adopt the maximum operator for the location, so that, the output for the kth template R^{HC2_k} is calculated as

$$R_k^{\mathrm{HC2}} = \max_j R_{k,j}^{HS2}. \tag{16}$$

As the result, when we prepare K templates to describe input data set, we can obtain K dimension vector for an input image. This local template-matching and maximum integration might be effective for the deformation of the image caused by the object translation.

4. Computer Simulation & Results

In the evaluation of the H-HOG using compute simulation, we adopt following parameters. In the experiment using conventional HOG, we prepare several sizes for the cell, which is the unit description for the histogram described as eq.(6). These sizes are $\{5 \times 5, 10 \times 10, 15 \times 15, 20 \times 20, 25 \times 25\}$ [pixels^2]. The quantization parameter is fixed as $Q = 9$, and the block size is also fixed as 3×3[cells^2]. In the following, we denote the HOG descriptor that have $n \times n$ cells size as HOG$_n$ for convenience.

In the H-HOG experiment, the model has multi-resolution representation in the pre-$HS2$ layer. We prepare several cell sizes for the multi-resolution representation, that are $\{5 \times 5, 10 \times 10, 15 \times 15, 20 \times 20, 25 \times 25\}$ cells. For template choosing in the $HS2$-layer denoted as eq.(12), template block size are randomly chosen from following candidates: $\{1 \times 2, 2 \times 4, 3 \times 6\}$ [blocks^2].

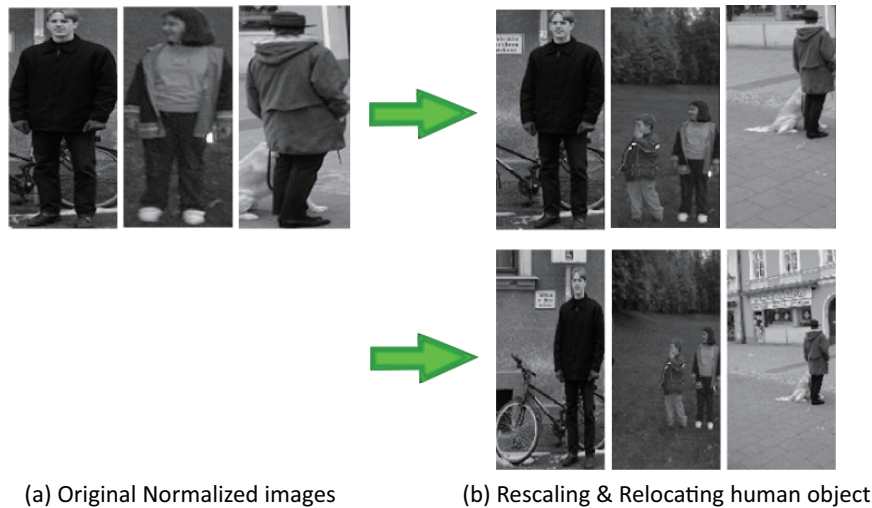For the H-HOG experiments, we focus to the feature extraction ability in $HS2$ layer,

(a) Original Normalized images   (b) Rescaling & Relocating human object

Fig. 4 Examples of the human images. The left shows those in the original INRIA data set provided by Dalal 100×200[pixels^2].The right top shows those of the data-set 1, which is re-cropped into 180×400[pixesl^2] and re-scaled into 100×200 [pixels^2]. The right bottom shows those of the data-set 2, which is re-cropped into 300×600 [pixels^2] and re-scaled into 100×200[pixels^2].

so that, we prepare the following three types of H-HOGs. One is H-HOG with applying template selection described in section3.3.1, and we denote it as "H-HOG$_{sel}$". We choose $U = 0.8$ for the threshold in eq.(13) to discard similar templates, which is determined experimentally. The second type is applying the non-linear modulation denoted as eq.(10) into the H-HOG$_{sel}$. We described this type as "H-HOG$_{sel}^{non}$". The modulation parameters, which are threshold h and emphasize factor d, are experimentally determined as $h = 50$ and $d = 1.1$ respectively. The last one is not applying these modifications, and we denote it without any suffixes as "H-HOG".

4.1 Dataset: INRIA Person Detection

We evaluate the performance of the H-HOG for a person detection problem using a modified INRIA person data-set.The conventional INRIA person data-set is for the evaluation of the classification accuracy of the conventional HOG. In the conventional data-set provided by Dalal[1], human objects are segmented and normalized in the 64×128 [pixels^2]. Fig. 4 left shows several examples of the normalized human image in the same manner of Dalal except image size, which is 100×200[pixels^2]. Roughly speaking,

the parts of the human such that head, body, and legs in the normalized images looks located similar position, and each size of human looks same size even though the target is child or adult.

In this study, we re-crop the human images from the original image database for evaluation of the scale and location invariance. The data-set 1 is cropped as 180×400[pixels^2] from the original image, and normalized in 100×200 [pixels^2], which includes human object at random position for positive samples. The data-set 2 is also cropped as 300×600[pixels^2] from the original image, and normalized in 100×200[pixels^2]. The locations and sizes of human objects are assigned more random rather than those of the data-set 1 for positive samples. Fig.4 right top shows corresponding examples in the data-set 1, and the bottom shows the data-set 2. The difference between data set 1 and 2 is human objects sizes and locations. The larger cropping images, which is included in the data-set 2, have more flexibility rather that that of the smaller set. Thus, the data-set 2 is considered to be the most difficult for human detection in our prepared data-sets, since the size and location is further from the normalized images provided by Dalal[1].

For obtaining templates in the $HS2$ layer of the H-HOG, we prepare 200 inputs images for template creation. The contents of these images are 100 positive examples that involve human object, and 100 negatives that does not involve. For creating proper templates, we use scale and location normalized images for positive samples in the same manner with Dalal such like images in fig.4 left. In our simulation, our algorithm select 20 templates for each training image, so that $HS2$ layer would have $4,000$ templates at a maximum, and similar templates would be discarded in the H-HOG$_{sel}$ and H-HOG$_{sel}^{non}$ by applying templates selection described in section 3.3.1.

For evaluation, we use both data-set 1 and 2. In each data-set, 80 images are used for training of the classifier SVM for both the conventional HOG and the H-HOG descriptions. The number of positive images and negatives are 40 images equivalently. We apply the SVM provided from the OpenCV with default parameters.We also prepare another 600 patterns for each data-set in order to evaluate the classification accuracy of the conventional HOG and H-HOG. The positives and negatives are also equivalently included.

4.2 Detection performance for INRIA person Dataset

The table 1 shows the result of detection performance for data-set 1 and 2. The data-set 1 has small variations of human object locations in normalized 100×200[pixels^2]. Each column shows, detecting accuracy, input dimension size for classifier, spending

Table 1 Result for the data-set 1 and 2. First columns shows input dimension. From the 2nd to the 4th column shows the result for the data-set 1, and the 5th to the 7th shows that for the data-set 2.

	# Description	Data-set 1			Data-set 2		
	Dimensions	Accuracy[%]	time[sec]		Accuracy[%]	time[sec]	
			Learn	Test		Learn	Test
HOG_5	55,404	88.2	140.80	2.92	71.5	140.82	2.89
HOG_{10}	11,664	88.2	27.88	0.55	75.3	27.77	0.55
HOG_{25}	972	78.3	2.31	0.04	80.5	2.28	0.03
H-HOG	4,000	86.3	8.73	0.12	83.3	8.54	0.11
$H\text{-}HOG_{sel}$	2,431	86.8	5.34	0.07	83.0	5.25	0.06
$H\text{-}HOG_{sel}^{non}$	3,306	88.7	7.04	0.08	86.5	7.01	0.10

time for learning, and for testing respectively. We investigate these performances over the computer which has following specification: OS: Ubuntu 10.04 LTS, CPU: Xeon E5530 2.4 [GHz], Memory: 24 [GBytes]. In the conventional HOG, the small size of cell denoted as HOG_5 shows better performance on the accuracy rather than the other HOGs, while the size of input dimension for classifier, and spending time is larger than those of others. Our H-HOG shows also good performance, while the input dimensions is only $4,000$ at a maximum. Moreover, H-HOG introduced non-linear modulation and template selection, which is denoted as $H\text{-}HOG_{sel}^{non}$, show the best accuracy result for the data-set 1. In the result of the data-set 2, the HOG_5, which shows the best detecting accuracy in the data-set 1, indicates the worst accuracy.So that, the conventional HOG requires designing for adjusting to the detecting object scales. On the contrary, H-HOG also shows the good performance against to the conventional HOGs.

5. Conclusion

In this study, we propose a hierarchical extension of HOG model, and evaluate the performance about classification. In the conventional HOG models, the HOG_5, HOG_{10} shows a good result for the small image data-set, however, the dimension of input vector for the classifiers become over $10,000$ dimensions. Generally, the large dimension classification brings several difficulties for classification, that is called 'the curse of dimensionality'. On the contrary, our H-HOG model can control the dimension, which is the number of the templates selected from multi-resolution representation of the HOG description. We also demonstrate that the performance of our model is as good as that of the HOG_5 even in the small dimensions, which is under $4,000$ dimensions, of the input vector for classifier.

Moreover, our model shows the performance of the classification for the data-set 2 is the best result. The data-set 2 have larger flexibility for the location and size of the human. The performances of the classification of conventional HOG models are just affected to these flexibilities. The smaller cell size becomes, the worse the classification result becomes. On the contrary, our H-HOG model integrates several resolution size and show robustness for these flexibilities. Thus, we consider these hierarchical extension is effective for the generic object recognition for the real world.

References

1) Dalal, N. and Triggs, B.: Histograms of Oriented Gradients for Human Detection, *International Conference on Computer Vision & Pattern Recognition*, Vol.2, pp.886–893 (2005).
2) Yanai, K.: The Current State and Future Directions on Generic Object Recognition, *IPSJ Transactions on Computer Vision and Image Media*, Vol.48, pp.1–24 (2007). (in Japanese).
3) Fujiyoshi, H.: Gradient-Based Feature Extraction : SIFT and HOG, Technical Report 2007-CVIM-160, IPSJ (2007). (in Japanese).
4) Felleman, D.J. and van Essen, D.C.: Distributed hierarchical processing in primate cerebral cortex, *Cerebral Cortex*, Vol.1, pp.1–47 (1991).
5) Hubel, D.H. and Wiesel, T.: Sequence Regularity and Geometry of Orientation Columns in Monkey Striate Cortex, *J.Comp.Neurol.*, Vol.158, pp.267–293 (1974).
6) Fukushima, K.: Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position, *Biological Cybernetics*, Vol.36, No.4, pp.193–202 (1980).
7) Shouno, H.: Recent Studies around the Neocognitron, *ICONIP 2007, Revised Selected Papers, Part I* Lecture Notes in Computer Science, Vol.4984, Springer, pp.1061–1070 (2007).
8) LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-Based Learning Applied to Document Recognition, *Proceedings of the IEEE*, Vol.86, No.11, pp.2278–2324 (1998).
9) Mutch, J. and Lowe, D.: Multiclass Object Recognition with Sparse, Localized Features, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol.1, pp.11–18 (2006).